Revisiting Three Comparisons of Unobserved Conditional Invariance Techniques for the Detection of Differential Item Functioning

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Quintin Ulysses Adrian Love

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ernest C. Davenport, Jr, Geoffrey M Maruyama

December, 2014

Acknowledgements

First and foremost, I would like to thank God! Without the Lord's favor and grace, I have nothing. Hallelujah.

Second, I would like to thank my advisors, Ernest C. Davenport, Jr. and Geoffrey M. Maruyama. Thanks for believing in me when others in the program did not. More importantly, thanks for supporting me and helping me develop into a scholar.

Third, I would like to thank my committee members, Lesa C. Clarkson and Mark L. Davison. My dissertation topic and interest in this problem would not exist without our various conversations. I have learned so much from you both over the past few years.

Fourth, I would like to thank the supervisor of my internship, Nancy Walters. Thanks for supporting me as I completed my coursework and requirements toward graduation. To be specific, thanks for recognizing me as a student employee versus an employee that was also a student.

Finally, I would like to thank my CGC family, Venoreen Browne-Boatswain and Michelle Kuhl, and all of my colleagues in the Quantitative Methods in Education program. Without the opportunities and support provided by CGC and the University of Minnesota and the grace of God, I would not have accomplished this feat!

Dedication

I would like to dedicate this dissertation to my wife, Shantel. Without your love and support, I would not have survived graduate school.

Abstract

Within social science research, data are often collected using a measurement instrument that produces ordered-categorical data. When comparing scores created from a measurement instrument across subpopulations, measurement invariance must be a tenable assumption. Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) are two unobserved conditional approaches to assessing measurement invariance. Within the research literature, there are three often cited simulation studies that compare the two unobserved conditional invariance techniques. Because the research design of the three studies varied greatly, the results of the studies are contradictory and not comparable. In this simulation study, the true positive (TP) and false positive (FP) rates of the IRT and CFA approaches to assessing measurement invariance are evaluated under four manipulated factors: (a) source of Differential Item Functioning (DIF), (b) size of DIF, (c) sample size, and (d) baseline model. The parameters used for the data generation came from a five-item unidimensional scale with four ordered-categories (i.e., Likert-scale). The results suggest that the IRT model using a free-baseline is the most precise model. Additionally, regardless of the model chosen, a free-baseline model is most favorable across all conditions of source of DIF, size of DIF, and sample size. Finally, the TP and FP rates of the studied models vary as a function of source of DIF, size of DIF, sample size, and baseline model. The significance of these results for social science research is discussed.

Table of Contents

List of Tables

CHAPTER ONE

INTRODUCTION

Within the social sciences, many constructs of interest cannot be directly observed (Bollen, 1989). For instance, within educational research, a student's algebraic ability cannot be directly measured, and thus must be approximated using a measurement instrument. During the development phase of the measurement instrument, it is assumed that the measurement instrument produces scores that are accurate approximations of the student's level of algebraic ability. It is further assumed that the measurement instrument will produce scores that are accurate approximations for all individuals of the population of interest. Put simply, it is assumed that the measurement instrument is not biased towards individuals based on an unrelated grouping factor (Meredith, 1993). Subsequent to the development phase of the measurement instrument, these assumptions are tested.

There are two ways to assess whether a measurement instrument is producing estimates that are not systematically accurate or biased (Millsap & Everson, 1993). First, there are observed conditional invariance techniques for assessing measurement bias. Falling within this category are procedures such as the Mantel-Haenszel χ^2 method (Holland & Thayer, 1988; Mantel & Haenszel, 1959, as cited by Millsap & Everson, 1993), standardization approaches (Dorans, 1989), and logistic regression methods (Swaminathan & Rogers, 1990).

There are also unobserved conditional invariance techniques for assessing measurement bias (Millsap & Everson, 1993). Falling within this category are procedures based on Factor Analysis (FA) and Item Response Theory (IRT). Theoretically, the

various approaches for assessing measurement bias should lead to the same items being flagged as exhibiting Differential Item Functioning (DIF). However, given the unobserved conditional approaches to assessing measurement bias, the research literature suggests that this is often not the case (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004, April).

Within the measurement invariance literature, there are three simulation studies that are often cited by applied researchers: Kim and Yoon (2011), Meade and Lautenschlager (2004), and Stark, Chernyshenko, and Drasgow (2006). These studies are influential because they focused on the similarities and differences of FA and IRT based approaches to assessing DIF; however, they are limited in that they are not comparable. As a consequence, there are still many questions that remain.

1.1 Statement of the problem

The literature is limited in that most studies compared the linear Confirmatory Factor Analysis (LCFA) model to an IRT model (e.g., Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006). Even though research suggests that the LCFA model is appropriate for 5-point Likert scaled data under conditions of normal skewness and kurtosis (Babakus, Ferguson, & Jöreskog, 1987; Muthén & Kaplan, 1985), the results may not extend to multiple groups CFA (MG-CFA) models (Lubke & Muthén, 2004; Temme, 2006). For instance, Stark et al. (2006) found that the LCFA model is appropriate for detecting DIF using 5-point Likert scaled data, whereas Meade and Lautenschlager (2004) found that the LCFA model has low power for detecting DIF. It is important to note that the DIF

generation methods differed across studies. As a consequence, the results are dissimilar (Elosua, 2011).

Kim and Yoon (2011) furthered the literature by comparing the categorical CFA (CCFA) model to the 2PL and GR IRT models. However, because the LCFA model was not included in the study, the CCFA results are not comparable to Meade and Lautenschlager (2004) or Stark et al. (2006). In addition, because threshold DIF is simulated on all thresholds, which is the same DIF generation method used by Stark et al. (2006), the results cannot be compared to Meade and Lautenschlager's (2004) study, which generated threshold-parameter DIF on specific thresholds.

Taking all three of the studies into consideration, because Meade and Lautenschlager (2004) used a different DIF generation method from Stark et al. (2006) and Kim and Yoon (2011) and Kim and Yoon (2011) used a different CFA model from the CFA model used by Meade and Lautenschlager (2004) and Stark et al. (2006), the results across studies are not comparable (Elosua, 2011). It is argued that the studies are not comparable because of the differences in the methods and designs across studies. Thus, research is needed that compares the LCFA, CCFA, and IRT models using both methods of DIF generation within a single study.

Finally, similar to many simulation studies involving Structural Equation Modeling (SEM; Paxton, Curran, Bollen, Kirby, & Chen, 2001), the literature is limited because the population parameters simulated may not adequately represent empirical data (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004). In other words, external validity is often a limitation of simulation studies, as parameter values are often selected

arbitrarily. To my knowledge, Stark et al. (2006) is the only study that used population parameters that were based on empirical data. Additional research is needed using population parameters based on empirical data.

1.2 Purpose of the study

The purpose of this dissertation study is to answer the following research questions:

1. Using the LCFA model, the CCFA model, and the Graded Response (GR) IRT model, which model is most precise at detecting DIF given unidimensional ordered-categorical data?
    a. Do the findings vary as a function of the source of DIF?
    b. Do the findings vary as a function of the size of DIF?
    c. Do the findings vary as a function of the type of baseline model used for identification?
    d. Do the findings vary as a function of sample size?

1.3 Significance of the study

There are multiple ways that the findings of this study can be significant to research within the social sciences and psychometrics. First, while several simulation studies have been conducted to examine the conditions under which the performance of FA and IRT based approaches to assessing DIF are similar and/or different (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006), there are differences in the methodology of the studies, with the result that optimal approaches are not discernible from the existing research. The current study incorporates features of the prior three

studies so it can unify findings that compare unobserved conditional invariance techniques to assessing DIF. For instance, the current study includes the LCFA model, the CCFA model, and the GR model. As a consequence, the similarities and differences of the models under various data conditions can be compared directly.

Second, the current study is significant because of its possible impact for applied researchers using the CFA model to analyze five-point Likert-scaled data. The need for simulation research that studies the unobserved conditional invariance approaches to assessing DIF is based on early empirical research (Raju et al., 2002; Reise et al., 1993). Even though the literature has progressed to the point where the original questions proposed by Reise, Widaman, and Pugh (1993), and Raju, Laffitte, and Byrne (2002) are being studied using simulation research, the methods and designs of the simulation studies have led to additional complexities. As an example, given a five-point Likert-scaled measurement instrument, should factorial invariance be studied using the LCFA or CCFA model? Research suggests that studying the factorial invariance of ordinal data using the LCFA model may lead to distorted factor structures and, as a consequence, produce biased results (Lubke & Muthèn, 2004; Temme, 2006). To my knowledge, Lubke and Muthèn (2004) is the only study to test this issue. Therefore, this study also contributes to the literature that compares the performance of the LCFA to the CCFA model given five-point Likert-scaled data.

CHAPTER TWO

LITERATURE REVIEW

2.1 Measurement

McDonald (1999) argues that "the purpose of measurement is to quantify an attribute" (p. 55). Within the social sciences, many attributes or constructs of interest (e.g., attitudes, behaviors, traits) are unobservable or latent (Bollen, 1989). Assuming that the construct of interest is an abstract concept, measurement is "the process by which a concept is linked to one or more latent variables, and these are linked to observed variables" (Bollen, 1989, p. 180). A latent variable or factor is defined as an "unobservable variable that influences more than one observed [variable] and that accounts for the correlations among these observed [variables]" (Brown, 2006, p. 13). An observed or manifest variable is defined as a variable that can be directly observed (e.g., item scores) and serves as an indicator of a latent variable.

Measurement models are mathematical expressions that operationalize the relationships between latent and observed variables, and lead to quantification of the construct(s) of interest (Bollen, 1989). Measurement instruments are needed to collect the manifest variables required for the measurement models. If a purpose of the measurement instrument is to make valid comparisons of groups (i.e., subpopulations), then measurement invariance is an assumption that must be tested and found to be tenable (Millsap, 2011).

2.2 Measurement Invariance

Measurement invariance exists when the conditional distribution of an observed variable, given a value of the latent trait or common factor, is independent of group membership (Meredith, 1993; Meredith & Teresi, 2006; Millsap, 2011). Measurement invariance can be expressed as

$$P_g(\boldsymbol{x}|\boldsymbol{\xi}) = P(\boldsymbol{x}|\boldsymbol{\xi}) \quad (2.1)$$

where $P_g(\boldsymbol{x}|\boldsymbol{\xi})$ is the conditional probability function for $\boldsymbol{x}$ given $\boldsymbol{\xi}$ in group g, g (g = 1, 2,…G) represents a grouping variable, $\boldsymbol{x}$ ($\boldsymbol{x} = x_1, x_2,\ldots x_i$) is a random vector of scores on the i (i = 1, 2,…j, I) manifest variables, and $\boldsymbol{\xi}$ ($\boldsymbol{\xi} = \xi_1, \xi_2,\ldots \xi_m$) is a random vector of scores on m (m = 1, 2,…M; $M < I$) common factors (Millsap, 2011). Equation 2.1 provides a general probabilistic structure that can be applied to any measurement model. Depending on the measurement model fit to the data (e.g., IRT models, FA models), measurement invariance can be examined at the item and/or test level. Given the definition of measurement invariance, measurement bias can be defined (Millsap, 2011; Millsap & Everson, 1993).

2.3 Measurement Bias

Measurement bias is defined as a violation of measurement invariance (Millsap, 2011; Millsap & Everson, 1993). Measurement bias exists when the measurement instrument provides a systematically inaccurate estimate of latent ability for subgroups within a population (e.g., gender, race/ethnicity). As a consequence, measurement bias can lead to decisions based on invalid observations (Ackerman, 1992; Drasgow, 1982; Osterlind & Everson, 2009). The focus of this paper is on measurement bias based on

unobserved conditional invariance, which is tested directly using a measurement model that relates the observed score to the unobserved latent trait score (Millsap & Everson, 1993). The two measurement frameworks of interest are IRT models and FA models. The following sections will describe measurement bias within each measurement framework, the measurement models of each measurement framework, and tests of measurement bias within each framework. Following the discussion on measurement bias, studies that compared measurement bias methods across measurement frameworks are reviewed.

2.4 Factorial Invariance

If a purpose of the instrument is to make valid comparisons of subgroups, then measurement invariance is required (Meredith, 1993; Meredith & Teresi, 2006; Millsap, 2011). Within the FA literature, measurement invariance is termed factorial invariance (Gregorich, 2006; Meredith, 1993; Meredith & Teresi, 2006; Millsap, 2011). Assuming a common factor model, factorial invariance exists when the factor structure underlying the data is the same across subgroups.

2.5 Common Factor Model

The common factor model can be expressed as

$$\boldsymbol{x} = \boldsymbol{\tau} + \boldsymbol{\Lambda\xi} + \boldsymbol{\delta} \tag{2.2}$$

where $\boldsymbol{\Lambda}$ (commonly called the factor pattern matrix) is a $i \times m$ matrix of factor loadings (λ_{im}) relating the i manifest variables to the m common factors, $\boldsymbol{\tau}$ ($\boldsymbol{\tau} = \tau_1, \tau_2, \ldots \tau_i$) is a vector of intercepts for the i manifest variables, and $\boldsymbol{\delta}$ ($\boldsymbol{\delta} = \delta_1, \delta_2, \ldots \delta_i$) is a random vector of scores on the i unique factors (Jöreskog, 1969; MacCallum, 2009; Millsap, 2011). It is important to note that the unique factor, δ_i, is the sum of the item specific factor, s_i, which

influences only item *i*, and random measurement error, e_i. Under the common factor model, it is assumed that

$$Cov(\boldsymbol{\xi}, \boldsymbol{\delta}) = 0, \tag{2.3a}$$

$$Cov(\boldsymbol{\xi}) = \boldsymbol{\Phi}, \tag{2.3b}$$

and

$$Cov(\boldsymbol{\delta}) = \boldsymbol{\Theta}, \tag{2.3c}$$

where $\boldsymbol{\Phi}$ is the $m \times m$ matrix of common factor covariances, and $\boldsymbol{\Theta}$ is a $i \times i$ diagonal matrix of unique factor variances (Jöreskog, 1969; MacCallum, 2009). Without loss of generality

$$E(\boldsymbol{x}) = \boldsymbol{\tau} = 0, \tag{2.4a}$$

$$E(\boldsymbol{\xi}) = 0, \tag{2.4b}$$

and

$$E(\boldsymbol{\delta}) = 0, \tag{2.4c}$$

which leads to the covariance structure of $\boldsymbol{x}$

$$\boldsymbol{\Sigma} = E(\boldsymbol{x}\boldsymbol{x}') = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2.5}$$

where $\boldsymbol{\Sigma}$ is the $i \times i$ population covariance matrix for the manifest variables (Jöreskog, 1969; MacCallum, 2009). Equation 2.5 defines the variances and covariances of the manifest variables as a function of the model parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$, and thus represents a theory about the structure of the population covariances among the manifest variables (Bollen, 1989; MacCallum, 2009).

Depending on the researcher's theory, the common factor model can be fit in an exploratory or confirmatory manner (Bollen, 1989; Brown, 2006; Long, 1983;

MacCallum, 2009). Generally speaking, if a researcher is lacking a prior theory, then the parameters of the common factor model are estimated using an unrestricted or Exploratory Factor Analysis (EFA); conversely, if the researcher has a theory a priori, then the parameters of the common factor model are estimated using a restricted or CFA (Bollen, 1989; Brown, 2006; MacCallum, 2009). Early studies on factorial invariance utilized the EFA model; however, now that the CFA model has been developed, and it is able to test a wide range of invariance hypotheses, the CFA model is used most often in present research (Millsap, 2011). Consequently, only factorial invariance using CFA models is further discussed. The LCFA model is discussed in greater detail (Bollen, 1989; Brown, 2006; Jöreskog, 1969; MacCallum, 2009), followed by a discussion of the CFA model appropriate for ordered-categorical data (i.e., CCFA; Christoffersson, 1975; Muthén, 1978, 1984). Subsequent to the discussion on the LCFA and CCFA models, tests of factorial invariance are discussed (Millsap, 2011; Muthén & Christoffersson, 1981; Vandenberg & Lance, 2000).

2.6 Confirmatory Factor Analysis

The purpose of FA is to account for the covariances among a set of manifest variables using a smaller set of common factors (i.e., latent variables; Bollen, 1989; Brown, 2006; Long, 1983; MacCallum, 2009). Contrary to EFA, CFA requires that the researcher have a theory that guides specification of the parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}$. Assuming that the model's specifications lead to model identification, the model parameters are estimated, producing a unique solution. Given the estimated model parameters ($\widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{\Phi}}, \widehat{\boldsymbol{\Theta}}$) and Equation 2.5, an implied covariance structure is created

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Lambda}}' + \widehat{\boldsymbol{\theta}}, \quad (2.6)$$

and subsequently compared to the sample covariance structure ($\boldsymbol{S}$) of observed variables, which is a $i \times i$ matrix. Model parameters are estimated to find values that produce an implied covariance structure ($\widehat{\boldsymbol{\Sigma}}$) that reproduces the sample covariance structure ($\boldsymbol{S}$). The accuracy of the model's ability to reproduce the sample covariance structure is assessed using a fitting or discrepancy function, $F(\boldsymbol{S},\widehat{\boldsymbol{\Sigma}})$. The fitting function equals zero if the model's parameters lead to an exact replication of the sample covariance matrix.

The fitting function depends upon the estimation method (e.g., unweighted least squares, maximum likelihood; Bollen, 1989; Long, 1983; MacCallum, 2009). The choice of estimation method depends on the nature of the data (Bollen, 1989; Long, 1983). Each estimation method has a set of assumptions about the data, and violation of the model's assumptions may bias results (e.g., Babakus, Ferguson, & Jöreskog, 1987; DiStefano, 2002). Within the factorial invariance literature, maximum likelihood (ML) is the estimation method most often used in applied (e.g., Davis-Kean & Sexton, 2009; Gregorich, 2006) and simulation (e.g., Meade & Lautenschlager, 2004; Stark et al., 2006) studies, and thus is the only estimator for continuous measures discussed in this paper. Subsequent to the discussion of the ML estimation method, the CCFA model (Christoffersson, 1975; Muthén, 1978; 1984) is discussed.

2.6.1 Maximum likelihood.

Based on the ML estimation method, the discrepancy function can be expressed as

$$F_{ML} = tr\left(\boldsymbol{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right) + \left[ln\left|\widehat{\boldsymbol{\Sigma}}\right| - ln|\boldsymbol{S}|\right] - i, \quad (2.7)$$

where $tr(\cdot)$ is the trace of a matrix, namely, the sum of the diagonal elements, and $ln|\cdot|$ is the natural log of the determinant of a matrix (Bollen, 1989; Brown, 2006; Jöreskog, 1969; Long, 1983). It is assumed that the observations are independent, the sample is sufficiently large, the model is correctly specified, and the data are continuous and multivariate normally distributed. Estimates are produced using an iterative procedure that minimizes the ML discrepancy function. The minimum of the ML discrepancy function value can be used to calculate a chi-square (χ^2) test statistic to assess the null hypothesis that the specified model fits exactly in the population (Brown, 2006; Long, 1983; Millsap, 2011), and is calculated as

$$\chi^2 = F_{ML}(N-1), \tag{2.8}$$

where N is the sample size. Assuming that the model is correctly specified, the test statistic is χ^2 distributed with degrees of freedom equal to $[i * (i + 1)] / 2$ minus the number of free model parameters (Bollen, 1989). If the assumptions of the model and estimation method are tenable, then the ML estimates are asymptotically unbiased (i.e., the expected value of the parameter estimates approach the population parameters in large samples), asymptotically efficient (i.e., the variance of the sampling distribution of the ML estimators is at a minimum in large samples), and consistent (i.e., the parameter estimates converge to the population parameters as sample size increases; e.g., Bollen, 1989; Brown, 2006; Finney & DiStefano, 2006; Jöreskog, 1969; Long, 1983; MacCallum, 2009).

2.7 Confirmatory Factor Analysis Model for Ordered-Categorical Data

The previous sections described a factor model and estimation method (i.e., ML) appropriate for continuous measures (Bollen, 1989; Brown, 2006; Long, 1983; MacCallum, 2009). However, in applied research in the social sciences, data tend to be ordinal (DiStefano, 2002; Flora & Curran, 2004; Lubke & Muthén, 2004). In spite of the ordinal nature of the data, applied researchers assess factorial invariance using a LCFA model with ML estimation (Lubke & Muthén, 2004; Sass, 2011). As previously mentioned, the CFA model using ML estimation assumes the following: (a) independent observations, (b) large sample sizes, (c) a correctly specified model, and (d) continuous data that are multivariate normally distributed (e.g., Finney & DiStefano, 2006; Maruyama, 1998). When the observed data are on an ordinal scale, the assumption of multivariate normally distributed data is violated, and as a consequence, ML may produce biased results (Babakus et al., 1987; DiStefano, 2002; Finney & DiStefano, 2006; Lubke & Muthén, 2004).

Even though there are multiple estimators that can handle ordered-categorical data, such as the Asymptotic Distribution Free estimator (ADF; Browne, 1984) and Generalized Least Squares (GLS; Christofferson, 1975; Muthén, 1978), research suggests that the Robust Weighted Least Squares (RWLS) approach is optimal (Flora & Curran, 2004). This is due to the large sample size needed for stable estimates using the ADF and GLS estimation methods (Flora & Curran, 2004; Millsap, 2011). Moreover, as the number of indicators increases, the minimum sample size required also increases. This is not the case using the WLS approach (Muthén, 1983, 1984) to testing factorial invariance

described by Muthén and Christofferson (1981), and thus is the only estimation method further discussed.

2.7.1 Latent response variates

Prior to discussion of the WLS estimation method (Muthén, 1983, 1984), the common factor model must be modified to accommodate the ordinal nature of the data (Christofferson, 1975; Muthén, 1978, 1983, 1984). The common factor model expressed in Equation 2.2 defines the common factors in $\boldsymbol{\xi}$ as continuous variables linearly related to the manifest variables, and thus assumes that the manifest variables are also continuous (e.g., MacCallum, 2009; Wirth & Edwards, 2007). Given ordered-categorical data, this is not a tenable assumption. As a result, the common factor model and estimation procedure must be modified to accommodate the nature of the data. Christofferson (1975) and Muthén (1978, 1984) describe the use of a threshold model that relates the ordinal manifest variable (x_i) to a latent response variate (x_i^*), which follows a standard normal distribution with a mean of 0 and variance of 1, expressed as

$$x_i = c \text{ if } v_{ic} < x_i^* < v_{ic+1}, \tag{2.9}$$

where v_{ic} ($v_{ic} = v_{i0}, v_{i1}, v_{i2}, \ldots, v_{ic}$) are the c thresholds that relate the latent response variate to the response category k ($k = 1, \ldots, c + 1$). It is further assumed that $v_{i0} = -\infty$ and $v_{ic+1} = \infty$. The thresholds of the latent response variates can be estimated using the following equation

$$v_{ic} = \phi^{-1}\left(\sum_{k=1}^{c} \frac{N_k}{N}\right) \tag{2.10}$$

where ϕ^{-1} is the inverse of the standard normal distribution function, and N_k is the number of subjects who selected category k (Bollen, 1989; Finney & DiStefano, 2006;

Olsson, 1979; Wirth & Edwards, 2007). Thus, the threshold model estimates threshold parameters that mark the point on the continuous latent response scale that separates one manifest discrete response option from the adjacent response option (e.g., Yes/No). Given the threshold model in Equation 2.10, Equation 2.2 can be modified to model the factor structure underlying the latent response variates

$$\boldsymbol{x}^* = \boldsymbol{\tau} + \boldsymbol{\Lambda\xi} + \boldsymbol{\delta} \tag{2.11}$$

where $\boldsymbol{x}^*$ ($\boldsymbol{x}^* = x_i^* = x_1^*, x_2^*, \ldots x_i^*$) is a random vector of scores on the *i* latent response variables. Because a single latent response variate is assumed to follow a univariate normal distribution, it is further assumed that a pair of latent response variates follow a bivariate normal distribution. As a consequence, the correlation structure of the latent response variables can be expressed as

$$\boldsymbol{P}^* = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}, \tag{2.12}$$

where $\boldsymbol{P}^*$ is the $i \times i$ polychoric correlation matrix (Bollen, 1989; Finney & DiStefano, 2006; Wirth & Edwards, 2007).

2.7.2 Polychoric correlations

As previously discussed, the factor structure is modeling the underlying latent response variates (Bollen, 1989; Finney & DiStefano, 2006; Wirth & Edwards, 2007). Subsequently, the correlation and covariance matrices calculated from the observed ordinal data are not of interest. Instead, the correlations of interest are those between the latent response variates called tetrachoric/polychoric correlations. Assuming polytomous data, there are two ways described by Olsson (1979) to estimate the thresholds and polychoric correlations. The first option is to estimate the thresholds and polychoric

correlations simultaneously using ML. The second option is to first estimate the thresholds as the inverse of the standard normal distribution, evaluated at the cumulative marginal proportions of the contingency table (see Latent response variate section). Subsequent to estimation of the threshold parameters, the ML estimate of the population correlation coefficient is computed, conditional on the estimated thresholds. Because the difference between the estimates of the two procedures is minimal and the two-step procedure takes less time to compute, the two-step procedure may be preferred. The following section discusses estimation methods appropriate for the models shown in Equations 2.9-2.12.

2.7.3 Weighted least squares

Given polychoric correlations, the WLS estimator is the only method that produces correct standard errors and test statistics (Bollen, 1989, p. 443). The WLS fitting function can be expressed as

$$F_{WLS} = [\boldsymbol{r} - \boldsymbol{\rho}]' \boldsymbol{W}^{-1} [\boldsymbol{r} - \boldsymbol{\rho}], \tag{2.13}$$

where $\boldsymbol{r}$ represents a vector of unique elements of the $i \times i$ sample polychoric correlation matrix and thresholds, $\boldsymbol{\rho}$ represents a vector of unique elements of the $i \times i$ model implied polychoric correlation matrix and thresholds, and $\boldsymbol{W}$ represents a consistent estimator of the asymptotic covariance matrix of the elements of $\boldsymbol{r}$ (Bollen, 1989; Muthén, 1983; Muthén, du Toit, & Spisic, 1997; Wirth & Edwards, 2007). Depending on the number of items, the size of $\boldsymbol{W}$ may be very large (Brown, 2006; Muthén, 1993; Muthén et al., 1997; Wirth & Edwards, 2007). To ensure that the estimate of the $\boldsymbol{W}$ matrix is correct and that

the $\boldsymbol{W}$ matrix can be inverted, the sample size must be larger than the number of elements in $\boldsymbol{W}$.

As a result, Muthén (1993) and Muthén, du Toit, and Spisic (1997) developed RWLS by setting all off-diagonal elements of the $\boldsymbol{W}$ to zero, so that it is a diagonal matrix, and subsequently easier to invert when obtaining standard errors of parameter estimates (Brown, 2006; Muthén et al., 1997; Wirth & Edwards, 2007). Because of the reduction in information (i.e., $\boldsymbol{W}$ is a diagonal matrix), the $\boldsymbol{W}$ is no longer the optimal weight matrix, and subsequently estimates obtained are not statistically efficient (Muthén et al., 1997; Wirth & Edwards, 2007). Because of the lack of efficiency, the standard errors and test statistics are biased. Consequently, inspired by Satorra (1992, as cited by Muthén et al., 1997), Muthén (1993) argued for the use of a mean and mean and variance adjustment that corrects the chi-square test statistic and standard error of the parameter estimates (see Muthén, 1993 and Muthén et al., 1997 for a detailed discussion). It is important to note that the calculation of the robust test statistics avoids inversion of the full $\boldsymbol{W}$ matrix, and therefore, may be used with sample sizes as small as 100 cases (Flora & Curran, 2004).

2.8 Multiple Groups Confirmatory Factor Analysis

The following section discusses factorial invariance (Gregorich, 2006; Meredith, 1993; Meredith & Teresi, 2006; Millsap, 2011). Prior to discussing the parameters of interest when assessing factorial invariance, the common factor model must be extended to model multiple groups (i.e., multiple group confirmatory factor analysis; MG-CFA). According to Jöreskog (1971) and Millsap (2011), MG-CFA can be expressed as

$$\boldsymbol{x}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\xi}_g + \boldsymbol{\delta}_g. \quad (2.14)$$

In addition, it is no longer assumed that the latent and manifest variables are deviation scores, and thus,

$$E(\boldsymbol{\xi}_g) = \boldsymbol{k}_g, \quad (2.15a)$$

and

$$E(\boldsymbol{x}_g) = \boldsymbol{\mu}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\kappa}_g, \quad (2.15b)$$

which allows for each group to have a mean structure for the common factor and observed scores ($\boldsymbol{\kappa}_g$ and $\boldsymbol{\mu}_g$, respectively).

Using ordered-categorical data, the mean structure does not include the factor intercept matrix (Millsap, 2011). Given the multiple thresholds needed to be modeled using ordered-categorical data, the mean structure can be expressed as

$$\boldsymbol{\mu}_g = \boldsymbol{\Lambda}\boldsymbol{\kappa}_g, \boldsymbol{\nu}_g. \quad (2.16)$$

The covariance structure for g groups ($\boldsymbol{\Sigma}_g$) can be expressed as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\,\boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g. \quad (2.17)$$

The ML discrepancy function for MG-CFA is

$$F_{ML} = \sum_{g=1}^{G}\left(\frac{N_g}{N}\right)F_{ML_g} \quad (2.18)$$

where N_g is the total sample size for group g, and F_{ML_g} is the ML discrepancy function for group g. Assuming CCFA, the threshold model is extended to accommodate the multiple groups

$$x_{ig} = c \text{ if } v_{icg} < x_{ig}^{*} < v_{ic+1g}, \quad (2.19)$$

and the WLS discrepancy function for MG-CFA is

$$F_{WLS} = \sum_{g=1}^{G} F_{WLS_g} \quad (2.20)$$

(Millsap, 2011).

2.8.1 Tests of factorial invariance

Factorial invariance is assessed using a series of nested models and the chi-square test statistic (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). The following section describes the tests of factorial invariance in detail. Assuming a MG-CFA model using ML estimation, the parameters of interest are the $\boldsymbol{\tau_g}$, $\boldsymbol{\Lambda}_g$, and $\boldsymbol{\Theta}_g$ matrices (Jöreskog, 1971; Millsap, 2011). Assuming a MG-CFA model using WLS, the parameters of interest are the $\boldsymbol{\nu_g}$, $\boldsymbol{\Lambda}_g$, and $\boldsymbol{\Theta}_g$ matrices (Millsap, 2011; Muthén & Christoffersson, 1981). Subsequent to the discussion of the tests of factorial invariance, partial invariance is discussed (Byrne, Shavelson, & Muthén, 1989).

2.8.1.1 Invariance of covariance matrices

The initial step is to test the null hypothesis of invariant population covariance matrices ($\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$; Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). Failure to reject the null hypothesis suggests that the equality of population covariance matrices is plausible, and thus there is no need for further invariance testing (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). On the contrary, rejection of the null hypothesis suggests that some degree of measurement non-invariance exists (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). Differences in the covariance matrices suggest that the inter-relationships of the manifest variables are different for the different groups. This could happen if for a group of academic tests we had dropouts versus graduates, and the graduates took the tests more seriously, and thus the scores were more

reliable for them. In such a case, we may expect to have higher inter-correlations of the measures for the graduates.

Byrne (1998, as cited by Raju et al., 2002) argues that the test of covariance matrices invariance is not needed because the test results may lead to contradictory results. For instance, the null hypothesis of invariant covariance matrices may be rejected even though subsequent invariance tests of specific parameter matrices may suggest that measurement invariance is a tenable assumption (e.g., Meade & Lautenschlager, 2004). As a consequence, many applied researchers studying measurement invariance do not include the test of invariant covariance matrices (see Vandenberg & Lance, 2000 for a review of applied measurement invariance research).

2.8.1.2 Configural invariance

Factorial invariance is assessed using a series of nested models and the χ^2 test statistic (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). The first test of invariance is configural invariance. Configural invariance is assessed by setting the same pattern of fixed and free elements in the $\boldsymbol{\Lambda}_g$ parameter matrices. Configural invariance can be expressed as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g. \quad (2.21)$$

Because the free elements are not constrained to be equal across groups, configural invariance implies that similar, but not identical, latent variables are conceptualized across groups. Failure to reject the null hypothesis implies that the same number of latent variables underlies the data and that the latent variables correspond to the same set of manifest variables. Rejection of the null hypothesis suggests that different latent variables

underlie the data across groups, and subsequent tests of factorial invariance are not justified. As an example, assuming item responses on a motivation scale and a dropout group and a graduate group, rejecting the null hypothesis of configural invariance may occur when the clusters of observed item responses are not identical across groups suggesting that the construct of motivation is being measured differently across groups (Vandenberg & Lance, 2000).

2.8.1.3 Metric invariance

Metric invariance constrains the factor pattern matrices to be fully invariant (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). Metric invariance can be expressed as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Phi}_g \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_g. \tag{2.22}$$

Notice that the g subscript is no longer attached to the $\boldsymbol{\Lambda}$ matrix. Failure to reject the null hypothesis suggests that the relationships between the latent and manifest variables are the same across groups. If the null hypothesis is rejected, then the relationships between the latent and manifest variables vary across groups. Returning to the example previously discussed, this may occur when a cluster of items on the motivation scale does not have the same meaning for dropouts and graduates (Gregorich, 2006).

2.8.1.4 Scalar invariance

Scalar invariance is the first test that involves the mean structure of the data (Millsap, 2011). Using continuous data, scalar invariance constrains the vector of intercepts in the mean structure to equality, and can be expressed as

$$\boldsymbol{\mu}_g = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}_g. \tag{2.23}$$

Scalar invariance is also termed strong factorial invariance (Meredith, 1993). Scalar invariance implies that differences in means on the manifest variables are due to differences in means on the latent variables (Millsap, 2011; Vandenberg & Lance, 2000). Failure to reject the null hypothesis suggests that the factor intercepts are the same across groups, whereas rejection of the null hypothesis suggests that the factor intercepts vary across groups. Rejecting the null hypothesis of scalar invariance may occur if, for example, the graduate group is more likely to exhibit an acquiescence response style bias than the dropout group on the motivation scale (Gregorich, 2006).

Using ordered-categorical data, the test analogous to scalar invariance constrains the item thresholds ($\boldsymbol{\nu_g}$) to be invariant (Millsap, 2011), and can be expressed as

$$\boldsymbol{\mu}_g = \boldsymbol{\Lambda\kappa}_g, \boldsymbol{\nu}. \tag{2.24}$$

Assuming that the latent response variates are multivariate normally distributed, the test of threshold invariance is a test of strong invariance. Threshold invariance suggests that the amount of the common factor needed to select a specific response category is invariant across groups, and furthermore, that differences in means on the manifest variables are due to differences in means on the latent variables. Failure to reject the null hypothesis suggests that the item thresholds are the same across groups, whereas rejection of the null hypothesis suggests that the item thresholds vary across groups. Similar to scalar invariance using continuous responses, rejecting the null hypothesis may occur when the acquiescence response bias exists within a subgroup of the population (Gregorich, 2006). For instance, the graduate group is more likely to agree with

individual items on the motivation scale then the dropout group. The primary difference is whether the item responses are continuous or categorical.

2.8.1.5 Invariance of unique factor covariance matrices

Assuming that the focus of the study is on the items of the measurement instrument, the final test of factorial invariance is the test of invariant unique factor covariances (Millsap, 2011). Meredith (1993) used the term strict factorial invariance. According to Millsap (2011), invariance of the unique factor covariance matrices can be expressed as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Phi}_g \boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \quad (2.25)$$

Under strict invariance, group differences in covariances among observed variables are attributable only to differences in covariances among latent variables (Millsap, 2011; Vandenberg & Lance, 2000). Assuming that the common factor variances are the same, a test of invariant unique factors is analogous to a test for invariant reliabilities across groups (Vandenberg & Lance, 2000). Failure to reject the null hypothesis suggests that the matrix of item unique variance is the same across groups, whereas rejection of the null hypothesis suggests that the matrix of item unique variance varies across groups (Millsap, 2011; Vandenberg & Lance, 2000). Moreover, upon verifying the assumption of invariant unique factor covariance matrices, the measurement instrument is assumed to be free of measurement bias (Millsap, 2011). As previously discussed (see Invariance of covariance matrices section), if the graduate group took the series of tests more seriously than the dropout group, the null hypothesis of the strict invariance test may be rejected due to the differences in reliability.

2.8.1.6 Partial invariance

The invariance tests previously mentioned (e.g., configural invariance, metric invariance, scalar invariance) assumed that the elements of the entire parameter matrices met the assumption of invariance (i.e., full invariance; Byrne et al., 1989; Millsap, 2011; Millsap & Kwok, 2004). Partial invariance exists when at least one element of the parameter matrix of interest violates that assumption of invariance. That is, after rejecting the null hypothesis of one of the aforementioned tests of factorial invariance (e.g., configural invariance, metric invariance), the source of the invariance may be found by relaxing the constraints on individual items within a parameter matrix. Subsequently, partial invariance allows for comparisons across groups even if full measurement invariance is untenable. A limitation of partial factorial invariance is that there are no prescribed or minimum set of conditions to achieve partial invariance (Millsap & Kwok, 2004). For example, given the motivation scale and the groups of dropouts and graduates, partial invariance may occur when the relationship between one or more of the items on the motivation scale is not identical across the dropout and graduates groups.

2.9 Differential Item Functioning

Within the IRT framework, measurement bias may be present when DIF exists (Millsap, 2011; Millsap & Everson, 1993). DIF refers to differences in item functioning after conditioning on the ability intended to be measured by the item (Dorans & Holland, 1993, as cited in Raju et al., 2002; Osterlind & Everson, 2009). That is, DIF exists when the parameters of item response functions (IRF) are non-invariant across groups (Osterlind & Everson, 2009). DIF is a required, but not sufficient condition for item bias

(Millsap, 2011; Millsap & Everson, 1993; Osterlind & Everson, 2009). Consequently, a DIF analysis is required to ensure valid interpretation of test scores across groups. Prior to discussion of the various methods of DIF using IRT, the IRT models of interest are discussed.

2.10 Item Response Theory Models

IRT is within the family of latent variable models (Mislevy, 1986; Thissen & Steinberg, 1986), and models observed data at the item-level (Birnbaum, 1968; De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). The first IRT models focused on dichotomous data (Lord & Novick, 1968). IRT assumes that the observed item responses are a function of an individual's common factor and the item properties (Birnbaum, 1968; De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). For instance, the 3-paramter dichotomous IRT model assumes that the item's discrimination, difficulty, and probability of guessing interact with an individual's latent ability to produce an observed item response (Birnbaum, 1968).

The regression of the item response on the underlying common factor leads to the IRF or item characteristic curve (ICC; Birnbaum, 1968; De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). Assuming dichotomous data, as the level of the common factor (i.e., latent trait) increases, the proportion of individuals endorsing the item increases; however, it is apparent that the relationship is not linear (De Ayala, 2009; McDonald, 1999). Given the monotonic relationship between observed scores and the proportion of individuals endorsing an item, IRT models use a non-linear function to model the observed item response data (Birnbaum, 1968; De Ayala, 2009; Hambleton &

Swaminathan, 1985; Lord & Novick, 1968; McDonald, 1999). Because of the ogival pattern, the cumulative normal or logistic distribution can be used to model the data. A discussion on the assumptions of the IRT models, which apply to both the normal ogive and logistic IRT models, follows.

The common assumptions of IRT focus on the dimensionality of the latent space, local independence, and functional form of the ICC (De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). The assumption of dimensionality states that the response data are a manifestation of one or more common factors. Once all of the common factors that underlie the data are accounted for in the model, the latent space is complete. In practice the assumption is relaxed, and it is assumed that there is (are) a dominant factor(s) that influences item responses (Hambleton & Swaminathan, 1985). Relaxing the assumption of dimensionality allows for it to be further assumed that the latent space is complete. As a consequence, assuming an instrument designed to measure mathematics ability, an incorrect response is due solely to a lack of mathematics ability.

The assumption of local independence states that, conditional on the common factor(s), the responses to an item are statistically independent of the responses to any other item on the instrument (De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968; McDonald, 1999). Put simply, all item responses should be uncorrelated at a fixed level of the common factor. Violating the assumption of local independence implies that some examinees have higher expected test scores than others of the same ability level (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). As a

consequence, additional common factors would be needed to account for the examinee's performance, which suggests that the test is multidimensional.

The functional form assumption states that the data follow the specified IRT model (De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). McDonald (1999) argues that there are three properties required for a function representing the conditional probability of endorsing an item. That is, the function should (a) "be bounded above by unity, and below by zero", (b) "be smooth and monotone-increasing", and (c) "approach horizontal asymptotes at each extreme value of [the common factor]" (McDonald, 1999, p. 250). The functional form of IRT models is that of an ogive, and is represented by the trace line of the ICC. The number of parameters used to describe the ICC distinguishes the family of IRT models (Osterlind & Everson, 2009). The following sections introduce the IRT models commonly used in applied research, and describe the functional form of each model in greater detail.

2.11 IRT Models for Dichotomous Data

The normal ogive model was the first item response model, and posits the normal cumulative distribution as a response function for item *i* (Birnbaum, 1968; De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). Because the lower asymptote is bounded by zero and the upper asymptote is bounded by one, the normal ogive model may be used as a probabilistic model (De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968; McDonald, 1999). The normal ogive model is expressed as the integral

$$P(x_i = 1|\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad (2.26)$$

where $P(x_i = 1|\theta)$ is the probability that a randomly selected examinee with common factor θ endorses item i, a_i is the discrimination parameter for item i, b_i is the difficulty parameter for item i, and z is a normal deviate from a distribution with mean b_i and standard deviation $1/a_i$ (De Ayala, 2009; Hambleton & Swaminathan, 1985). The a_i is proportional to the slope of the tangent at the point on the common factor scale where the probability of endorsing the item is 50%, and the b_i is the point on the common factor scale where the probability of endorsing the item is 50% (De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968). It is important to note that the normal ogive model does not assume that examinees of extremely low ability will select the correct response option by chance (i.e., guessing does not occur; Lord & Novick, 1968, p. 383); however, assuming multiple choice items, Birnbaum (1968) extended the model to account for guessing (p. 404). A limitation of the normal ogive model is that the integration is not easily performed (Birnbaum, 1968; Hambleton & Swaminathan, 1985). Consequently, logistic models are used in most applications of IRT (Thissen & Steinberg, 1986).

Birnbaum (1968) introduced the item response model using the logistic distribution. The decision to model the data using a logistic distribution has several advantages over the normal ogive distribution (Birnbaum, 1968). For instance, "the logistic model is more 'mathematically tractable' than the normal ogive model because the latter involves an integration while the former is an explicit function of item and ability parameters" (Hambleton & Swaminathan, 1985, p. 37). Consequently, the logistic IRT models are "mathematically convenient" (Thissen & Steinberg, 1986). In addition,

with the use of a scaling factor D = 1.7, the normal and logistic distributions nearly coincide (Birnbaum, 1986; Hambleton & Swaminathan, 1985).

IRT models are distinguished by the number of parameters estimated about the items (Osterlind & Everson, 2009). The first logistic model for dichotomous data discussed is the three-parameter (3PL) model, and can be expressed as

$$P(x_i = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad (2.27)$$

where c_i is the lower asymptote of the IRF and is known as the guessing parameter (Birnbaum, 1968). Assuming that the c_i is greater than 0, the b_i is located at the point of the common factor scale where the slope of the ICC is a maximum, and is no longer on the common factor scale where the probability of endorsing the item is 50% (Hambleton & Swaminathan, 1985). Instead, the probability of endorsing the item is $(1 + c_i)/2$. In addition, the slope of the curve at b_i equals .425 $a_i(1 - c_i)$.

Similar to the normal ogive model, the two-parameter logistic (2PL) model does not assume that the probability of endorsing an item is a function of guessing (Birnbaum, 1968; Lord & Novick, 1968). The 2PL model can be expressed as

$$P(x_i = 1|\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad (2.28)$$

The interpretation of the parameters of the normal ogive and 2PL models are the same (Birnbaum, 1968; De Ayala, 2009; Hambleton & Swaminathan, 1985; Lord & Novick, 1968).

An even more restrictive model is the one-parameter logistic (1PL) model which assumes that the a_i are the same across all items on the instrument. The 1PL model can be expressed as

$$P(x_i = 1|\theta) = \frac{1}{1+ e^{-Da(\theta - b_i)}} \quad (2.29)$$

Given the 1PL model, items on a measurement instrument only vary by b_i. Assuming that the discrimination parameter of each item on the instrument is equal to 1, the 1PL model is known as the Rasch (1960, as cited by Birnbaum, 1968) model. Subsequently, the Rasch (1960, as cited by Birnbaum, 1968) model is a reparameterization of the 1PL model (Hambleton & Swaminathan, 1985).

2.12 IRT Models for Polytomous Data

There are many IRT models used for polytomous data (e.g., Rating Scale, Partial Credit Model, Generalized Partial Credit Model; Ostini & Nering, 2006; Thissen & Steinberg, 1986). The intent of this literature review is not to be exhaustive; therefore, the interested reader is referred to Ostini and Nering (2006) for a detailed account of polytomous IRT models. Based on a review of the measurement invariance literature, Samejima's (1969) GR model is the only polytomous model used in simulation research (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006). As a result, the GR model (Samejima, 1969) is the only polytomous IRT model discussed.

According to Millsap (2011), "the [GR model] views the response categories … as representing a series of steps" (p. 160). In other words, in order for an examinee to select a particular response category *c*, they must pass through all *c* - 1 categories that precede category *c* (Millsap, 2011; Ostini & Nering, 2006). The GR model specifies the

probability of an examinee endorsing category *c* as the difference between the probabilities of endorsing category *c* or above and endorsing category *c* + 1 or above, and can be expressed as

$$P(x_i = c|\theta) = \frac{1}{1+ e^{-Da_i(\theta - b_{ic})}} - \frac{1}{1+ e^{-Da_i(\theta - b_{ic+1})}}$$

$$= P_c^* - P_{c+1}^*, \quad (2.30)$$

where P_c^* represents the category boundary response function (CBRF), which is the trace line that describes the probability of responding in category *c* or higher, given the common factor, and b_{ic} is the boundary location for category *c* (Millsap, 2011; Ostini & Nering, 2006; Samejima, 1969). By definition, $P_0^* = 1$ and $P_k^* = 0$.

Given Equation 2.30, it is easy to see why Thissen and Steinberg (1986) classified the GR model as a difference model. In essence, the GR model is a two-parameter model that reduces the multiple response options into two sets of response options (De Ayala, 2009; Samejima, 1969; Thissen & Steinberg, 1986). Therefore, similar to the 2PL and 3PL models, the GR model does not have a sufficient statistic (Hambleton & Swaminathan, 1985; Millsap, 2011). Even though the GR model shown in Equation 2.30 follows a logistic distribution, it is important to note that the GR model can also be adapted to follow the normal ogive distribution (Samejima, 1969; Thissen & Steinberg, 1986).

2.12.1 Marginal maximum likelihood

The most commonly used parameter estimation methods in IRT are conditional maximum likelihood (CML), marginal maximum likelihood (MML), and Bayesian estimation (Millsap, 2011). Generally speaking, the most appropriate estimation method

depends on the IRT model. For instance, CML requires a sufficient statistic (Hambleton & Swaminathan, 1985; Millsap, 2011). Of the IRT models previously discussed, only the Rasch (1960, as cited by Birnbaum, 1968) model, which assumes that all item discrimination parameters are equal to one, has a sufficient statistic (Millsap, 2011). Subsequently, CML is only appropriate for the Rasch model. MML, however, does not require a sufficient statistic, and subsequently, is an appropriate estimation method for all of the IRT models previously discussed (Millsap, 2011; Thissen & Steinberg, 1986).

Without a sufficient statistic, both the item and person parameters must be simultaneously estimated; however, if the likelihood function can be expressed without any reference to the ability parameters, then the item parameters can be estimated (De Ayala, 2009; Hambleton & Swaminathan, 1985; Millsap, 2011). MML estimation "isolate[s] the item parameters by explicitly modeling the distribution of [the common factor], followed by integration over this distribution to yield a marginal likelihood for [the response vector]" (Millsap, 2011, p. 165). The marginal likelihood for response vector $\boldsymbol{x}$ can be expressed as

$$L(\boldsymbol{x}) = \int_{-\infty}^{\infty} P(\boldsymbol{x}|\theta) g(\theta) d\theta = \prod_{i=1}^{I} \int_{-\infty}^{\infty} P(x_i|\theta)\, g(\theta) d\theta \qquad (2.31)$$

where g(θ) is the population density function for θ specified from theory or given the empirical data and x_i represents a response vector from a random examinee in the population (De Ayala, 2009; Millsap, 2011). Given g(θ), the item parameters are estimated by maximizing the marginal likelihood function (Bock & Aitkin, 1981; De Ayala, 2009; Hambleton & Swaminathan, 1985; Millsap, 2011; Mislevy, 1986; Wirth & Edwards, 2007).

The integral in Equation 2.31 can be approximated by *q*-dimensional Gauss-Hermite quadrature, and can be expressed as

$$L(\boldsymbol{x}) = \sum_{q=1}^{Q} P(\boldsymbol{x}|Q_k)A(Q_k), \tag{2.32}$$

where the summation occurs over the q quadrature points, Q_k, and $A(Q_k)$ is the quadrature weight for each point Q_k (Bock & Aitkin, 1981; De Ayala, 2009; Mislevy, 1986; Wirth & Edwards, 2007). The Gauss-Hermite quadrature amounts to using q rectangles to approximate the area under the curve. Because the number of quadrature points corresponds to the number of rectangles used to approximate the area under the curve, increasing the number of quadrature points also increases the accuracy of the approximation. A negative consequence of increasing the number of quadrature points is an increase in the computation and complexity of the model (Cai, Yang, & Hansen, 2011). Cai, Yang, and Hansen (2011) argue that a minimum of 20 quadrature points are needed for a moderate degree of accuracy.

The MML estimates are found using an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) as described by Bock and Aitkin (1981; De Ayala, 2009; Millsap, 2011; Mislevy, 1986; Wirth & Edwards, 2007). The EM algorithm is an iterative procedure that proceeds in two steps. The first step is the expectation step (i.e., E-step), which involves using provisional estimates of the item parameters to obtain estimates for the expected number of endorsements for item *i* and the expected number of examinees along the regions of the *q* quadrature nodes, conditional on the data and item parameter estimates (Bock & Aitkin, 1981; De Ayala, 2009; Mislevy, 1986; Wirth & Edwards, 2007). The second step is the maximization step (i.e., M-step), which involves

taking the expected number of endorsements for item *i* and the expected number of examinees along the regions of the *q* quadrature nodes from the E-step as known, and obtaining new estimates of the item parameters by substituting the E-step estimates in the likelihood equations. The process continues until the desired convergence criterion is achieved. The parameter estimates obtained using MML are consistent and asymptotic normal (Hambleton & Swaminathan, 1985). The interested reader is referred to Bock and Aitkin (1981) for a detailed presentation of the formulas used during the E- and M-steps of the EM algorithm. The following sections will discuss approaches for assessing measurement bias using IRT models.

2.13 Tests of DIF

According to Millsap and Everson (1993), statistical approaches for assessing measurement bias can be categorized as either observed or unobserved conditional invariance models. Assuming unidimensionality, observed conditional invariance models condition on an observable random variable used to stratify the sample into subgroups (i.e., stratifying variable). Examples of observed conditional invariance models are the traditional χ^2 methods, the Mantel-Haenszel (MH) χ^2 method, standardization approaches, and logistic regression methods (Millsap & Everson, 1993; Osterlind & Everson, 2009). Unobserved conditional invariance models condition on a latent trait that is related to the observed items through a measurement model. Examples of unobserved conditional invariance models are IRT and FA (Millsap & Everson, 1993).

According to Osterlind and Everson (2009), "the aim of using IRT in DIF investigations is to determine whether an item assesses the underlying ability or

proficiency (θ) similarly for all groups taking the test in the portion of the ability continuum covered by the test item(s)" (p. 40). Within the IRT literature, there are numerous tests of DIF (e.g., Likelihood Ratio Test, Area-based measures; Millsap & Everson, 1993; Osterlind & Everson, 2009). Among the various IRT DIF methods, the likelihood ratio (LR; Thissen, Steinberg, & Wainer, 1988) test is most commonly used when comparing CFA and IRT (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006).

Assuming that the items are placed on the same scale across groups, the IRT LR test compares the likelihood when a studied item's parameters are constrained to be invariant across groups with the likelihood function when the parameters of the studied item are allowed to vary across groups (Millsap & Everson, 1993; Osterlind & Everson, 2009; Thissen et al., 1988). The LR can be expressed as

$$G^2 = -2 \ln\left[\frac{L(A)}{L(C)}\right], \tag{2.33}$$

where L(A) is the likelihood obtained when the studied item's parameter(s) are freely estimated across groups (i.e., augmented model) and L(C) is the likelihood obtained when the studied item has parameter(s) constrained to equality across groups (i.e., constrained model). The G^2 statistic is distributed approximately as a chi-square statistic with degrees of freedom equal to the number of constraints needed to derive the constrained model from the augmented model. The null hypothesis is that the item parameters are invariant across groups. The LR test can be applied to both dichotomous and polytomous data (Millsap & Everson, 1993).

The differential functioning of items and tests (DFIT) framework (Raju, van der Linden, & Fleer, 1995) is another approach to assessing DIF that has been used in comparisons of CFA and IRT (Raju et al., 2002). As a consequence, a brief review of the approach is warranted. Once again, assume that the item parameters are placed on a common scale across the reference and focal groups. The DFIT framework consists of three indices of differential functioning: (a) the differential test functioning (*DTF*) index, (b) the compensatory DIF (*CDIF*) index, and (c) the noncompensatory DIF (*NCDIF*) index (Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995).

Within the IRT context, an examinee's true score can be expressed as

$$T = \sum_{i=1}^{I} P(x_i|\theta), \tag{2.34}$$

According to the DFIT framework (Oshima et al., 1997; Raju et al., 1995), each examinee has a true score for being a member of the focal group [$T_F = \sum_{i=1}^{I} P_F(x_i|\theta)$] and reference group [$T_R = \sum_{i=1}^{I} P_R(x_i|\theta)$], where $P_F(x_i|\theta)$ and $P_R(x_i|\theta)$ represent the probability of success on item *i* for an examinee if they were in the focal group and reference group, respectively. If T_F equals T_R, then the examinee's true score is independent of group membership. The squared difference between T_F and T_R is the *DTF* index, and can be defined as

$$DTF = E_F(T_F - T_R)^2 = E_F D^2 = \sigma_D^2 + (\mu_{T\mathrm{F}} - \mu_{T\mathrm{R}})^2 = \sigma_D^2 + \mu_D^2, \tag{2.35}$$

where the expectation (*E*) is taken over the focal group in the current situation, but can be taken over either the focal or reference group, *D* equals the difference between true scores (T_F and T_R), and μ and σ refer to the mean and standard deviation, respectively (Oshima et al., 1997; Raju et al., 1995). If Equation 2.35 is rewritten as

$$DTF = E_F D^2 = E_F[\sum_{i=1}^{I}(d_i D)] = \sum_{i=1}^{I} E_F(d_i D) = \sum_{i=1}^{I}\left[Cov(d_i, D) + \mu_{d_i}\mu_D\right], \quad (2.36)$$

where d_i equals the difference between the probability of success for item *i* in the focal and reference group, $P_F(x_i|\theta) - P_R(x_i|\theta)$, $\sum_{i=1}^{I} d_i = D$, and $Cov(d_i, D)$ is the covariance between d_i and D, which reflects the correlated DIF between items such that

$$Cov(d_i, D) = \sigma_{d_i}^2 + \sum Cov(d_i, d_j), i \neq j, \quad (2.37)$$

then *CDIF*, an index of DIF at the item level, can be expressed as

$$CDIF_i = E_F(d_i D) = Cov(d_i, D) + \mu_{d_i}\mu_D, \quad (2.38)$$

such that $DTF = \sum_{i=1}^{I} CDIF_i$ (Oshima et al., 1997, p. 255). Because *DTF* is the sum of *CDIF*, an item with a positive *CDIF* may partially or fully cancel an item with a negative *CDIF*; hence, the term *CDIF* (Flowers et al., 1999; Oshima et al., 1997; Raju et al., 1995). Furthermore, according to Equations 2.37 and 2.38, the *CDIF* for item *i* includes correlated DIF between items *i* and *j* of the test (Oshima et al., 1997).

A second index of item level DIF is the *NCDIF*, which assumes that all items other than the one under study are free from DIF (Raju et al., 1995). If all items other than the studied item *i* do not exhibit DIF, then d_j equals zero for all $j \neq i$, and Equation 2.38 can be rewritten as

$$NCDIF_i = \sigma_{d_i}^2 + \mu_{d_i}^2 \quad (2.39)$$

The *NCDIF* is noncompensatory because its value for an item can only be non-negative and it assumes that all items other than the one under study are free from DIF. In addition, the *NCDIF* index is closely related to Lord's (1980, as cited by Raju et al., 1995) χ^2 test (Raju et al., 1995).

Raju et al. (1995) initially proposed significance tests based on the χ^2 distribution. However, because the χ^2 statistic is overly sensitive to large sample sizes, researchers began to identify cutoff values using Monte Carlo studies to decrease FP rates and power rates (e.g., Flowers et al., 1999; Raju et al., 1995). Oshima, Raju, and Nanda (2006) argue that applied researchers may not have the technical knowledge to develop cutoff values specific to their particular data sets, and, as a consequence, proposed a new method [i.e., item parameter replication (IPR) method] for deriving study-based cutoffs for assessing DIF in dichotomously scored items using the DFIT framework. Readers interested in the IPR method are referred to Oshima et al. (2006).

The DFIT framework has several advantages over other DIF methods. First, the DFIT framework can be applied using unidimensional data with dichotomous (Raju et al., 1995) and/or polytomous scoring (see Flowers, Oshima, & Raju, 1999). Second, the DFIT framework can be applied to multidimensional data (see Oshima, Raju, & Flowers, 1999; Flowers et al., 1999). Third, the DFIT framework can be applied at the test level by testing for DTF and the item level using *NCDIF* and *CDIF* indices (Flowers et al., 1999; Oshima, Raju, & Nanda, 2006). Fourth, the *CDIF* index does not assume that all items in the test other than the studied item are unbiased (Flowers et al., 1999). Finally, the DTF index provides the overall effect of eliminating an item from a test (Raju et al., 1995; Oshima et al., 1997).

Despite the advantages of the DFIT framework, the LR test is most often used in comparisons of CFA and IRT (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006). This is due to the similarity between measurement invariance using

CFA (i.e., χ^2) and the LR method (G^2; e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006). Nevertheless, future research using the DFIT framework in comparisons of measurement invariance using CFA versus IRT is warranted.

2.14 Comparisons between IRT and CFA

Reise et al. (1993) published the first article to compare measurement invariance using CFA and IRT. Specifically, the LCFA model is compared to the GR model using a 5-item negative affect measure called NA5. Respondents used a Likert-scale ranging from (1) not at all to (5) extremely to rate whether they considered themselves to be (a) nervous, (b) worried, (c) jittery, (d) tense, and (e) distressed. Data were collected from two samples of undergraduate students attending the University of Minnesota and the University of Nanjing Normal in China.

Reise et al. began by fitting a one-dimensional LCFA model to the five items using three different free-baseline models (i.e., minimal model constraints for identification) that were mathematically equivalent for illustrative purposes. Next, a full metric invariance model was fit to the data by constraining the factor pattern matrix to be equal across the samples of undergraduates. The results suggested that the full metric invariance was an untenable assumption. As a result, a partial metric invariance (Byrne et al., 1989) hypothesis was tested. The modification indices suggested that freeing the λ_{im} of the item distressed would lead to a significant decrease in the overall chi-square statistic (Reise et al., 1993). After freeing the λ_{im} of the item distressed across groups, there was a statistically significant improvement in model fit when compared to the full

invariance model; subsequently, partial invariance across samples of undergraduate students was tenable.

Because the goal of the paper was to identify similarities and differences across the CFA and IRT frameworks, similar analyses were conducted using the IRT approach to measurement invariance or DIF. Using the GR model, the baseline model was fit to the data using a concurrent calibration. To be specific, the data were treated as if 1,138 (540 Minnesota and 598 Nanjing) students responded to a 10-item test with items 1-5 missing for the Nanjing students and items 6-10 missing for the Minnesota students. Model fit was assessed using the IRT LR test (Thissen et al., 1988), whereas person fit was assessed using the Z_l statistic (Drasgow, Levine, & Williams, 1985), which is the standardized value of the likelihood of an individual's item response pattern given the IRF (Reise et al., 1993).

After fitting the baseline model, a full measurement invariance model was fit to the data by constraining all item parameters to equality (Reise et al., 1993). Because the full measurement invariance model led to a significant decrease in model fit, each item was tested for DIF separately. Similar to the results from the CFA analyses, the IRT analyses found that the item distressed was non-invariant across groups and the items nervous and tense were invariant across groups. Unlike the results from the CFA analyses, the IRT analyses suggest that the items worried and jittery also were functioning differently across groups. Consequently, a researcher's assessment of measurement invariance using the NA5 depends on the measurement framework used to test measurement invariance.

In similar fashion, Raju et al. (2002) compared factorial invariance using LCFA to the DFIT framework (Raju et al., 1995) using IRT. To be specific, the NCDIF index was used to assess DIF at the item level, and the DTF index was used to assess DIF at the subscale total score level. Data consisted of responses to a 10-item scale intended to measure satisfaction in work/assignment taken from the 1995 Armed Forces Sexual Harassment Survey (Edwards, Elig, Edwards, & Riemer, 1997). Data were collected from a sample of 1,000 Black and 1,000 White participants. Each of the 10 items on the scale has five response categories.

Beginning with the measurement invariance testing within the CFA framework, a free-baseline model was fit to the data, followed by a test of full metric invariance. The results suggest that full metric invariance is an untenable assumption; consequently, partial measurement invariance testing (Byrne et al., 1989) began by testing one item at a time using a nested model χ^2 analysis. Raju et al. (2002) found that only items 1 and 2 were noninvariant across racial groups.

Subsequent to the analysis using the CFA approach to factorial invariance, Raju et al. (2002) assessed measurement invariance using the DFIT framework. After calibrating the items within each group, the item parameters for the White group were placed on the same scale as the Black group and examined for DIF. Unlike the results from the CFA framework, only item 2 exhibited significant DIF using the NCDIF index. In addition, the DTF index suggested no signs of DIF at the subscale level. Based on the inconsistent results across frameworks (i.e., CFA suggesting DIF of items 1 and 2, whereas IRT suggesting DIF of only item 2), Raju et al. argued that "a comprehensive Monte Carlo

study, with some of the highlighted differences (e.g., dichotomous vs. polytomous data, sample size, the degree to which the underlying assumptions are met) between the two models as moderators, will be needed" (p. 527).

To my knowledge, Meade and Lautenschlager (2004) conducted the first simulation study comparing factorial invariance using CFA to DIF using IRT. The purpose of the study was to compare the ability of factorial invariance using LCFA and DIF using the IRT GR model. Measurement invariance was tested using the omnibus tests of invariant covariance matrices, and the tests of invariant factor pattern matrices, item intercepts, and factor variances. IRT DIF testing was based on the LR test (Thissen et al., 1988). The baseline model was identified by constraining a single referent item across groups and freely estimated all remaining items' parameters, factor variances, and factor means. Next, a series of models were fit to the data by constraining each item parameter separately. A statistically significant LR test statistic (i.e., G^2) suggests that the fit of the restrictive model is significantly poorer than that of the less restrictive model, and that the item may be exhibiting DIF.

A single scale composed of six Likert-scaled items with five response categories was simulated. First, data were created for group 1 (i.e., the reference group), then, based on a prescribed set of 15 DIF conditions, adjusted by subtracting 0.25 from the a_i and/or adding 0.40 to or subtracting 0.40 from the b_{ic} to create the data for group 2 (i.e., the focal group). The a_i estimates for the reference group were drawn from a random normal distribution with $\mu = 1.25$, $\sigma = 0.07$, whereas the b_{ic} estimates of the lowest CBRF of each item were sampled from a random normal distribution with $\mu = -1.7$, $\sigma = 0.45$. After

drawing a random value for the lowest CBRF of each item, the constants 1.2, 2.4, and 3.6 were added to the lowest CBRF to create the remaining three b_{ic}s needed to accurately represent an item with 5 response categories. The number of items exhibiting DIF was also manipulated by simulating either two of the six items exhibiting DIF or four of the six items exhibiting DIF. Finally, there were three sample sizes simulated for each group: 150, 500, and 1,000 with 100 samples simulated for each condition.

As briefly mentioned in the paragraph above, Meade and Lautenschlager (2004) created a total of 15 conditions of DIF. This is due to the fact that there were three different ways that DIF in the b_{ic} were simulated. That is, DIF in the b_{ic} was simulated by adding 0.40 to or subtracting 0.40 from: (a) the highest b_{ic} for each DIF item (e.g., strongly agree), (b) the two highest b_{ic}s for each DIF item (e.g., agree and strongly agree), and (c) the two extreme b_{ic}s for each DIF item (e.g., strongly disagree and strongly agree) Consequently, the items either did not exhibit DIF (one condition), only the a_i exhibited DIF (two conditions), only the b_{ic} exhibited DIF (six conditions), or both the a_i and b_{ic} exhibited DIF (six conditions).

Because the CFA tests of factor invariance test the entire parameter matrices, whereas LR tests test individual items, two sets of dependent variables were used (Meade & Lautenschlager, 2004). For comparisons made at the scale level, the dependent variable for the study was a statistic called AnyDIF, which recorded the number of times any of the six items was found to exhibit DIF. At the item level, the dependent variables for the study were the true positive (TP) rates, which record the number of items simulated to have DIF that were successfully detected as DIF items divided by the total

number of DIF items generated, and the false positive (FP) rates, which record the number of items flagged as DIF items divided by the total number of items simulated to not contain DIF.

Based on the results from the six conditions of DIF simulated in the b_{ic}, the CFA omnibus test of invariant observed covariance matrices was mostly incapable of detecting DIF unless the two highest b_{ic} or the two extreme b_{ic} exhibited DIF on four items with sample sizes of 500 and 1,000 (Meade & Lautenschlager, 2004). Furthermore, the subsequent tests of specific parameter matrices that follow a statistically significant omnibus test were unable to identify the specific source of the noninvariance. On the contrary, the LR test was largely effective at identifying the specific items contaminated with DIF, and increased in accuracy as the sample size and number of parameters exhibiting DIF increased. Similar results were found in the remaining conditions that manipulated DIF in the a_i and both the a_i and b_{ic}.

Finally, CFA TP and FP rates (.00-.20 and .00-.07, respectively) were very low overall and, surprisingly, the CFA TP rates were higher for sample sizes of 150 than 500 or 1,000. The IRT TP and FP rates (.15-1.00 and .02-.14, respectively), however, were higher for the larger samples (i.e., 500 and 1,000). As a result, Meade and Lautenschlager concluded that IRT using the LR test performs better in identifying DIF of the a_i and/or b_{ic} than does the test of measurement invariance using the CFA model.

Stark et al. (2006) compared CFA and IRT DIF detection methods. Specifically, factorial invariance using a mean and covariance structure (MACS) model was compared to the IRT LR test using the 2PL and GR models. A unidimensional scale consisting of

15 items was simulated using the linear common factor model and a threshold model that related the continuous measures to ordered-categories. The parameter values were obtained by analyzing real response data collected from the Illinois Supervisor Satisfaction Scale (ISSS; Chernyshenko, Stark, Crede, Wadlington, & Lee, 2003, as cited in Stark et al., 2006). Factors manipulated in the study were sample size (500 or 1,000), number of response categories (two or five), type of DIF (no DIF, λ_{im} DIF, v_{ic}s DIF, or DIF in both item parameters), amount of DIF simulated for focal group (small - λ_{im} = -.15, v_{ic}s = +.25 or large - λ_{im} = -.4, v_{ic}s = +.5), amount of impact [i.e. latent mean difference; 0.0 (i.e., common factor scores drawn from a standard normal distribution, N(0,1) for both the reference and focal groups) or 0.5 (i.e., common factor scores drawn from a standard normal distribution for the reference group and N(-0.5, 1) for the focal group)], type of baseline model (free-baseline model with only a referent item constrained to equality or constrained-baseline model with all items set equal across groups), and *p*-value for rejecting the null hypothesis of no DIF items [.05 or Bonferroni corrected (p-value/number of comparisons made)]. In the conditions of DIF, the data were simulated such that the 3rd, 7th, 11th, and 15th items exhibited DIF. Dependent variables used were the power or TP and the Type 1 error FP rates.

Stark et al. (2006) hypothesized that (a) the IRT LR method would better detect DIF when the data were dichotomous, (b) that the MACS method would perform better using small samples and polytomous data, and (c) that the IRT LR method would outperform MACS under conditions when both DIF and impact were present. Based on the results of the analyses using dichotomous data, Stark et al. found that under the

conditions of no DIF, the Type 1 error rate of the MACS analyses only exceeded the nominal value of .05 when the sample size was 1,000 and impact was present. Using the IRT LR method, the Type 1 error rate only exceeded the nominal value of .05 under the conditions with (a) a sample size of 1,000 and no impact (Type 1 error rate = .06) and (b) a sample size of 500, a free-baseline model, and impact (Type 1 error rate = .07). Using polytomous data and a constrained-baseline model, the Type 1 error rate of MACS analyses exceeded the nominal value of .05 under conditions (a) the sample was 500 and no impact and (b) impact was present, whereas using the free-baseline model, the Type 1 error rate only exceeded the nominal value in the condition where a MACS analysis was used with a sample size of 1,000 and impact present. On the contrary, the Type 1 error rate using the IRT method never exceeded the nominal value of .05.

Generally speaking, using the Bonferroni adjusted *p*-value, the Type 1 error rate did not exceed .01 in any of the simulated conditions for both the MACS and IRT DIF detection methods. If the Bonferroni adjusted *p*-values were not used, then the constrained-baseline condition led to inflated Type 1 error rates using both dichotomous (.06-.79) and polytomous (.00-.93) data across both the MACS and IRT LR DIF detection methods (Stark et al., 2006). On the contrary, the free-baseline model Type 1 error rates ranged from .00 to .12 across all manipulated conditions in the study. This suggests that a constrained-baseline model with anchor items exhibiting DIF may lead to higher FP rates.

DIF was easier to detect for v_{ic}s than for λ_{im} using both the CFA and IRT DIF detection methods. IRT DIF detection methods outperformed MACS DIF detection in

identifying DIF in large samples using dichotomous data. In contrast, using polytomous data, MACS DIF detection had higher power rates and similar Type 1 error rates as IRT DIF detection. It is important to note that these results are in contrast to those found by Meade and Lautenschlager (2004). To be specific, IRT outperformed CFA methods in DIF detection using polytomous data in Meade and Lautenschlager (2004), whereas Stark et al. (2006) found the opposite. It is also important to note that the threshold DIF conditions vary greatly between the two studies, which is argued as a limitation of the existing literature, and thus is a manipulated factor included in this dissertation. As a consequence, additional research is warranted. Finally, impact had little detrimental effect on the accuracy of both IRT and MACS DIF detection.

Kim and Yoon (2011) furthered the literature by comparing IRT to CCFA. Previous studies were limited because the "traditional" LCFA model was compared to IRT models (e.g., Meade & Lautenschlager, 2004; Stark et al., 2006). Given that ordered-categorical data are likely to violate the assumptions of LCFA, the comparisons between IRT and CFA were limited from the beginning (Reise et al., 1993). The comparisons were warranted, however, because applied researchers use either framework to test the hypothesis of measurement invariance (e.g., Oishi, 2006).

The purpose of Kim and Yoon's (2011) study was to compare the power to detect violations of measurement invariance using CCFA and IRT through a Monte Carlo study. Towards that end, the CCFA model with multiple groups was compared to the 2PL and GR IRT models. A balanced design with two groups of equal sample sizes: 100, 200, 500, and 1,000 were simulated. A unidimensional six item scale was simulated. Both

dichotomous and polytomous conditions were simulated. In conditions of polytomous data, five category ordered-response items were simulated. A single item was simulated to have DIF on the six item scale. Three different conditions of DIF were simulated: (a) noninvariance in the λ_{im}, (b) noninvariance in the v_{ic}(s), and (c) noninvariance in both the λ_{im} and v_{ic} (s). In addition, there were two conditions simulated for the size of DIF in the λ_{im} (i.e., small - .2 decrease and large - .4 decrease) and in the v_{ic}(s), (i.e., small - .3 decrease and large - .6 decrease) favoring the reference group. In all conditions, the unique variance of each item is set to .3 in both groups. Interestingly, the common factor mean and variance were simulated to be higher for the focal group (i.e., 0.5 and 1.3, respectively) than the reference group (i.e., 0.0 and 1.0, respectively).

The IRT LR test using a constrained-baseline model was compared to factorial invariance using a constrained-baseline model in CCFA (Kim & Yoon, 2011). The constrained-baseline model was identified by constraining the λ_{im} of each item to be equal across groups, setting the factor variance of the reference group to 1, and freely estimating all parameters in the focal group. The less-constrained (i.e., augmented) model relaxes the λ_{im} and v_{ic}(s) of an item to test for DIF. The critical *p*-value was adjusted using a Bonferroni correction, $p \leq .008$ (.05/6). Lastly, the performances of the root mean square error of approximation (RMSEA) and weighted root mean square residual (WRMR) statistics, which are alternative fit indices of CCFA, were also examined. Because these statistics are only appropriate for CFA, the results related to these fit indices will not be discussed.

Using the TP and FP rates as the dependent variables of the study, Kim and Yoon (2011) found that the TP rates in detecting DIF of the λ_{im} and v_{ic}(s) of IRT (.39-1.00) were similar to or higher than those of CCFA (.24-1.00) across all conditions using dichotomous data. However, when using polytomous data, the TP rates in detecting DIF of the λ_{im} and/or v_{ic}(s) of CCFA (.18-1.00) were the same as or higher than those of IRT (.08-1.00) across all manipulated conditions, which supports the findings of Stark et al. (2006). In addition, the FP rates were the same or higher for CCFA than IRT using dichotomous (.01-.86 and .01-.44, respectively) and polytomous (.01-.95 and .01-.29, respectively) data across all conditions.

Except for when the sample size is larger than 500 and the degree of DIF is large, λ_{im} DIF is difficult to detect for both methods using dichotomous data. DIF exhibited in the v_{ic} was difficult to detect when the sample size was small (i.e., 100, 200) and DIF was small for both methods. Both methods were able to detect large DIF in the v_{ic} or λ_{im} and v_{ic} across all sample size conditions (TP $\geq$ .93). CCFA performed as well as or outperformed IRT in detecting DIF when DIF was exhibited in the v_{ic} or λ_{im} only, whereas, if both the λ_{im} and v_{ic} exhibited DIF, then IRT performed as well as or outperformed CCFA in detecting DIF.

Using polytomous data, the TP rates of CCFA were either the same or higher than those of IRT across all conditions. In addition, except for the condition of large λ_{im} only DIF with a sample size of 500, the FP rates of CCFA were the same or higher than those of IRT. Even though CCFA outperformed IRT in detecting DIF exhibited in the λ_{im} only, both methods TP rates did not exceed .95 unless the size of DIF was large and the sample

was large (i.e., 500, 1,000). On the contrary, when detecting DIF of the v_{ic}s only and DIF of the λ_{im} and v_{ic}s, the TP rates of both measurement frameworks exceeded .95 when DIF was small and the sample size was large (i.e., 500, 1,000) and across all conditions of sample size when DIF was large for both methods.

Considering both the TP and FP rates, Kim and Yoon (2011) argue that IRT outperformed CCFA in DIF detection. That is, even though the TP rates of CCFA were slightly higher than those of IRT (i.e., .03 - 1.00 and .02 - 1.00, respectively) across most of the simulated conditions, the FP rates of CCFA were significantly higher than those of IRT (i.e., .01 - .95 and .01 - .44, respectively). Consequently, a researcher is as likely to correctly identify an item as exhibiting DIF and more likely to falsely identify an item as exhibiting DIF using CCFA than IRT.

2.15 Conclusion

In theory, if a measurement instrument is designed for comparisons across subpopulations, then the assumption of measurement invariance must be plausible. Millsap and colleagues (Millsap, 2011; Millsap & Everson, 1993) define measurement bias as a violation of measurement invariance. CFA and IRT are measurement models that can be used to assess measurement invariance; however, the conclusions drawn from each model's results may vary (Oishi, 2006; Raju et al., 2002; Reise et al., 1993).

Many simulation studies have compared the LCFA and IRT approaches to testing for measurement invariance or DIF using ordered-categorical data under various conditions (e.g., Forero & Maydeu-Olivares, 2009; Meade & Lautenschlager, 2004; Stark et al., 2006). Manipulated factors included sample size (e.g., Forero & Maydeu-Olivares,

2009, Kim & Yoon, 2011), type of DIF (e.g., Meade & Lautenschlager, 2004; Stark et al., 2006), and degree of impact (e.g., Stark et al., 2006). A review of the literature reveals that many studies comparing factorial invariance using CFA and DIF using IRT advantaged IRT because the simulated data often violated CFA model assumptions (Kim & Yoon, 2011; Lubke & Muthén, 2004; Reise et al., 1993).

For instance, several studies generated data using the GR model (Samejima, 1969) and compared the performance of CFA and IRT in identifying uniform (i.e., threshold) DIF (e.g., Meade & Lautenschlager, 2004; Stark et al., 2006). Because the LCFA model assumes continuous data and estimates a single item intercept versus multiple item thresholds, it was assumed that the IRT model would be better able to detect uniform DIF (Reise et al., 1993; Meade & Lautenschlager, 2004). Meade and Lautenschlager (2004) found that IRT was better able to detect b_{ic} DIF than CFA factorial invariance techniques, whereas Stark et al. (2006) found that CFA (i.e., mean and covariance structure analysis) outperformed IRT in identifying DIF using polytomous data. Because different models were used to generate the data and different strategies were used to simulate DIF, it is difficult to compare the results across studies. As a consequence, additional research is needed to further explore these contradictory results.

Kim and Yoon (2011) compared IRT and CCFA (Christofferson, 1975; Muthén, 1983; 1984; Muthén & Christoffersson, 1981). The categorical variable methodology assumes that underlying the ordered-categorical variables are latent response variates that are multivariate normally distributed (Christofferson, 1975; Kim & Yoon, 2011; Muthén, 1983; 1984; Muthén & Christoffersson, 1981). The latent response variables are

manifested as discrete scores with a set of thresholds and are estimated parameters in the CCFA model. As a result, polytomous data with multiple item thresholds can be correctly modeled using a CCFA model (Kim & Yoon, 2011; Muthén & Christoffersson, 1981; Reise et al., 1993). Despite the contributions made by Kim and Yoon (2011), there are a few limitations of their study that warrant additional research.

First, similar to Stark et al. (2006), Kim and Yoon (2011) simulated DIF on all v_{ic}s and found that CFA outperformed IRT in detecting v_{ic} DIF, which contradicts the results of Meade and Lautenschlager (2004). This may be due to the fact that Meade and Lautenschlager generated the data using an IRT model and simulated DIF on only one or two b_{ic}s. Meade and Lautenschlager assumed that certain groups may be less likely to select extreme item responses (e.g., Strongly Agree, Strongly Disagree), and, as a consequence, specific b_{ic}s may function differently across groups. Subsequently, additional research is needed that generates data using both the IRT and CFA model (Raju et al., 2002) and simulates DIF on all item thresholds (v_{ic}s and b_{ic}s) and only specific item thresholds (v_{ic}s and b_{ic}s), as outlined by Meade and Lautenschlager (2004).

Second, Kim and Yoon (2011) used a backward procedure with a constrained-baseline model to test DIF using a six-item scale with only one item exhibiting DIF in the conditions where DIF was present. Because research suggests that the CFA (i.e., MACS) model has more power to detect DIF when a constrained-baseline model approach is used as compared to a free-baseline model (Stark et al., 2006), research is needed that examines the performance of CCFA using a free-baseline model with minimal constraints needed for identification. Moreover, because Kim and Yoon (2011) simulated one DIF

item on the scale and used a constrained-baseline model, the constrained-baseline model was misspecified when assessing FP rates. Subsequently, additional research is needed that studies the performance of CCFA and IRT using a free-baseline model that is correctly specified.

Third, Kim and Yoon (2011) simulated equal sample sizes across groups. On the contrary, the focal and reference groups tend to be unbalanced in applied research (Woods, 2008, 2009). Consequently, research is needed that compares the power and FP rates of DIF detection using IRT to factorial invariance using CCFA under conditions of unbalanced designs. That is, research is needed to assess whether the power and FP rates of DIF vary as a function of the difference between the sample sizes of the reference and focal groups.

Finally, to my knowledge, missing from the literature are studies that study invariance of unique factor covariance matrices. For instance, Kim and Yoon (2011) simulated data using the CCFA model, and compared the ability of IRT and CCFA to detect DIF. Kim and Yoon (2011) simulated data such that the unique factor covariance matrices were invariant. Muthén and Christofferson (1981) argue that if the amount of unique factor variance varies across groups, then it is possible that the factor loadings are invariant while the item discrimination parameters are noninvariant. For this reason, additional research is needed that includes varying levels of the unique factors across groups as a manipulated factor.

CHAPTER THREE

METHODOLOGY

In chapter 2, I reviewed the literature on unobserved conditional invariance approaches to assessing measurement invariance (Millsap, 2011; Millsap & Everson, 1993). At the end of chapter 2, I identified various gaps in the literature. This dissertation study addressed two of those gaps through a simulation study. The purpose of this chapter is to discuss the methodology of the dissertation study. To be specific, the chapter will discuss the following: the (a) research design, (b) research questions, (c) unobserved conditional techniques for detecting measurement bias, (d) procedure, (e) data-generating mechanism, (f) data analysis, (g) manipulated factors, and (h) dependent variables of the study.

3.1 Research Design

The study used a quantitative research design. The research design was a simulation. Given that the study was a simulation study, it can be classified as experimental. The population parameters were based on data publicly released by the National Center for Education Statistics (NCES). They were obtained from a unidimensional five-item scale of school satisfaction collected during the 2007 wave of the National Household Education Surveys program (Hagedorn, O'Donnell, Smith, & Mulligan, 2008). Respondents were parents and guardians of elementary students.

The population parameters were based upon data from Love's (2012) empirical study. A scale composed of five items was simulated. It is important to note that having five items serve as indicators of a single latent factor is not unique to this study (e.g.,

Flora & Curran, 2004; Muthén & Kaplan, 1985; Reise et al., 1993). The factor loading parameters were based on an EFA model. Because the data are on a 4-point Likert scale, the procedure described by Stark et al. (2006, p. 1296) was used to create an additional response option. The five response options were used to create four thresholds by transforming the cumulative proportion of endorsement to a standard normal metric. The item labels, item factor loading parameters, and item thresholds parameters are shown in Table 3.1.

Similar to Kim and Yoon (2011), only a single item was manipulated to exhibit DIF. The factors and levels of each factor are shown in Table 3.2. Given the factors shown in Table 3.2 and the full factorial design of the study, there were a total of 54 DIF conditions. In addition, there were three conditions that focus on the FP rates with an unbiased anchor: (1) No DIF and a sample size of 250, (2) No DIF and a sample size of 500, and (c) No DIF and a sample size of 1,000. Therefore, there were a total of 57 conditions in the study. In an attempt to control for sampling error, a total of 1,000 replications were simulated for each condition. Finally, the dependent variables (i.e., TP and FP rates) were calculated as an average across all replications within each condition. A *p* value of .05 was considered statistically significant.

3.2 Research Questions

The purpose of this dissertation was to answer the following research question:

1. Using the LCFA model, the CCFA model, and the GR IRT model, which model is most precise at detecting measurement invariance of unidimensional ordered-categorical data?

Table 3.1

Population Parameters of Dissertation Study

Item	λ	v_1	v_2	v_3	v_4
Satisfied with the school that child attends?	0.93	-2.01	-1.46	-1.03	-0.42
Satisfied with teachers that child has this year?	0.74	-2.03	-1.54	-1.14	-0.59
Satisfied with the academic standards of school?	0.81	-1.96	-1.49	-1.06	-0.46
Satisfied with the order and discipline of school?	0.79	-1.83	-1.37	-1.02	-0.49
Satisfied with the way that school staff interacts with parents?	0.83	-1.94	-1.42	-0.98	-0.35

Note. Parameters are based on data collected by the National Household Education Survey: 2007 (Hagedorn, O'Donnell, Smith, & Mulligan, 2008). Respondents are parents and guardians of elementary school. Respondents responded to a 5-item, 4-point Likert scaled (i.e., Very dissatisfied, Somewhat dissatisfied, Somewhat satisfied, Very satisfied) school satisfaction scale. A fifth response option (i.e., neutral) was created using the procedure described by Stark, Chernyshenko, and Drasgow (2006). EFA model fit - χ^2 (5, N = 4,586) = 27.88, $p < .001$, Comparative Fit Index (CFI) = 1.00, Root Mean Square Error of Approximation (RMSEA) = 0.03, p = 1.00, Standardized Root Mean Square Residual (SRMR) = 0.02.

Table 3.2

Conditions and Manipulated Factors of Dissertation Study

	Manipulated Factors			
Condition	Source of DIF	Sample Size	Size of DIF	Baseline Model
1	-	250 500 1,000	-	Free Constrained
2	λ	250 500 1,000	Small Large	Free Constrained
3	v_1- v_4	250 500 1,000	Small Large	Free Constrained
4	v_4	250 500 1,000	Small Large	Free Constrained
5	v_3- v_4	250 500 1,000	Small Large	Free Constrained
6	v_1 and v_4	250 500 1,000	Small Large	Free Constrained
7	λ and v_1- v_4	250 500 1,000	Small Large	Free Constrained
8	λ and v_4	250 500 1,000	Small Large	Free Constrained
9	λ and v_3- v_4	250 500 1,000	Small Large	Free Constrained
10	λ and v_1 and v_4	250 500 1,000	Small Large	Free Constrained

Note. DIF stands for Differential Item Functioning. λ represents the factor loading given the Categorical Confirmatory Factor Analysis model. v_c represents the latent thresholds given the Categorical Confirmatory Factor Analysis model. In attempt to control for sampling error, 1,000 replications are simulated for each condition.

a. Do the findings vary as a function of sample size?

b. Do the findings vary as a function of the size and source of DIF?

c. Do the findings vary as a function of the type of baseline model used for identification?

3.3 Unobserved Conditional Techniques for Detecting Measurement Bias

3.3.1 Linear Confirmatory Factor Analysis

The MG-LCFA model can be expressed as

$$\boldsymbol{x}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\xi}_g + \boldsymbol{\delta}_g, \tag{3.1}$$

where $\boldsymbol{x}_g$ ($\boldsymbol{x}_g = x_{pig} = x_{p1g}, x_{p2g}, \ldots x_{pig}$) is a response vector for person p_g ($p_g = 1, 2, \ldots N_g$) of manifest variables i ($i = 1, 2, \ldots j, I$) in group g ($g = 1, \ldots G$), where the total sample size (N) equals the sum of N_g (i.e., the total sample size for group g), $\boldsymbol{\Lambda_g}$ is a $i \times m$ matrix of factor loadings (λ_{img}) relating the i manifest variables to the m common factors, $\boldsymbol{\tau_g}$ ($\boldsymbol{\tau_g} = \tau_{1g}, \tau_{2g}, \ldots \tau_{ig}$) is a vector of intercepts for the i manifest variables, and $\boldsymbol{\delta_g}$ ($\boldsymbol{\delta_g} = \delta_{1g}, \delta_{2g}, \ldots \delta_{ig}$) is a random vector of scores on the i unique factors for each group g (Jöreskog, 1969; MacCallum, 2009; Millsap, 2011). The covariance structure for group g can be expressed as

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g, \tag{3.2}$$

where $\boldsymbol{\Sigma}_g$ is the $i \times i$ covariance matrix, $\boldsymbol{\Phi_g}$ is the $m \times m$ matrix of common factor covariances, and $\boldsymbol{\Theta_g}$ is a $i \times i$ diagonal matrix of unique factor variances for group g (Millsap, 2011). Assuming a large sample size, multivariate normally distributed data,

correct model specification, and continuous measures, the LCFA model parameters are estimated using ML. The ML discrepancy function for MG-CFA is

$$F_{ML} = \sum_{g=1}^{G} \left(\frac{N_g}{N}\right) F_{ML_g}, \tag{3.3}$$

where F_{ML_g} is the ML discrepancy function for group g expressed as

$$F_{ML_g} = tr\left(\boldsymbol{S_g}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}^{-\mathbf{1}}\right) + \left[ln\left|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}^{-\mathbf{1}}\right| - ln\left|\boldsymbol{S_g}\right|\right] - i, \tag{3.4}$$

where $tr(\cdot)$ is the trace of a matrix, namely, the sum of the diagonal elements, $\widehat{\boldsymbol{\Sigma}}$ is the $i \times i$ model implied covariance matrix for the manifest variables, $\boldsymbol{S}$ is the $i \times i$ observed covariance matrix for the manifest variables, and $ln|\cdot|$ is the natural log of the determinant of a matrix (Bollen, 1989; Brown, 2006; Jöreskog, 1969; Long, 1983; MacCallum, 2009).

3.3.2 Categorical Confirmatory Factor Analysis

According to Millsap (2011), the group-specific threshold model relates the ordinal manifest variable (x_{ig}) for group g ($g = 1,\ldots, G$) to a latent response variate (x_{ig}^{*}), which follows a standard normal distribution with a mean of 0 and variance of 1, expressed as

$$x_{ig} = c \text{ if } v_{icg} < x_{ig}^{*} < v_{i(c+1)g}, \tag{3.5}$$

where v_{icg} ($v_{icg} = v_{i0g}, v_{i1g}, v_{i2g},\ldots, v_{i(c+1)g}$) are the thresholds that relate the latent response variate to the response category k ($k = 1,\ldots, c + 1$). It is further assumed that $v_{i0} = -\infty$ and $v_{ic+1} = \infty$. Given this assumption, there are a total of c thresholds that may vary.

Given the threshold model in Equation 3.5, the common factor model is modified to model the factor structure underlying the latent response variates

$$\boldsymbol{x}_g^* = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g\boldsymbol{\xi}_g + \boldsymbol{\delta}_g \tag{3.6}$$

where $\boldsymbol{x}_g^*(\boldsymbol{x}_g^* = x_{pig}^* = x_{p1g}^*, x_{p2g}^*, \ldots x_{pig}^*)$ is a random vector of scores on the *i* latent response variables for person p_g in group g (Jöreskog, 1969; MacCallum, 2009; Millsap, 2011). Because a single latent response variate is assumed to follow a univariate normal distribution, it is further assumed that a pair of latent response variates follow a bivariate normal distribution. As a consequence, the correlation structure of the latent response variables in each group can be expressed as

$$\boldsymbol{P}_g^* = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g, \tag{3.7}$$

where $\boldsymbol{P}_g^*$ is the $i \times i$ polychoric correlation matrix for group g (Millsap, 2011).

Given the polychoric correlations, the WLS estimator is the only method that produces correct standard errors, and test statistics (Bollen, 1989, p. 443). The WLS fitting function can be expressed as

$$F_{WLS} = \sum_{g=1}^{G}[\boldsymbol{r}_g - \boldsymbol{\rho}_g]' \boldsymbol{W}_g^{-1}[\boldsymbol{r}_g - \boldsymbol{\rho}_g], \tag{3.8}$$

where $\boldsymbol{r_g}$ represents a vector of unique elements of the $i \times i$ sample polychoric correlation matrix and thresholds for group g, $\boldsymbol{\rho_g}$ represents a vector of unique elements of the $i \times i$ model implied polychoric correlation matrix and thresholds for group g, and $\boldsymbol{W_g}$ represents a consistent estimator of the asymptotic covariance matrix of the elements of $\boldsymbol{r_g}$ for group g (Millsap, 2011).

Factorial invariance is assessed using a series of nested models and the chi-square test statistic (Jöreskog, 1971; Millsap, 2011; Vandenberg & Lance, 2000). Assuming a MG-CCFA model using WLS, the parameters of interest are those in the $\boldsymbol{v_g}$, $\boldsymbol{\Lambda}_g$, and $\boldsymbol{\Theta}_g$ matrices (Millsap, 2011; Muthén & Christoffersson, 1981). In an attempt to make the comparison of the CCFA comparable to IRT, the approach to measurement invariance described by Muthén and Muthén (1998-2010) is used, which is to free the item's discrimination and thresholds simultaneously. Muthén and Muthén (1998-2010) argue that DIF in either parameter is likely to lead to an item being discarded and/or revised. By taking this approach to assessing DIF, the tests of measurement invariance are comparable across measurement models.

3.3.3 Graded Response Item Response Theory Model

IRT models calculate the probability of an item response (or a pattern of item responses), conditional on the level of the common factor (De Ayala, 2009; Hambleton & Swaminathan, 1985). With the goal of making the analyses comparable across measurement models in mind, the model selected is the GR model.

The GR model can be expressed as

$$P(x_i = c|\theta) = \frac{1}{1+ e^{-Da_i(\theta - b_{ic})}} - \frac{1}{1+ e^{-Da_i(\theta - b_{ic+1})}}$$

$$= P_c^* - P_{c+1}^*, \quad (3.9)$$

where $P(x_i = c|\theta)$ represents the probability of an examinee endorsing category c given θ, P_c^* represents the CBRF, a_i is the discrimination parameter for item i, and b_{ic} is the

boundary location for category *c* (Millsap, 2011; Ostini & Nering, 2006; Samejima, 1969). By definition, $P_0^* = 1$ and $P_k^* = 0$.

According to Thissen (2001), the marginal likelihood for response vector $\boldsymbol{x}$ for two groups using an EM algorithm, as described by Bock and Aitkin (1981), can be expressed as

$$L(\boldsymbol{x}) = \prod_{g=1}^{2} \prod_{p_g=1}^{N_g} \int_{-\infty}^{\infty} P_g(\boldsymbol{x}|\theta) f_g(\theta) d\theta \tag{3.10}$$

where $f_g(\theta)$ is the population density function for θ within group g specified from theory or given the empirical data and the item parameters are embedded within $P_g(\boldsymbol{x}|\theta)$. Because the parameters for the focal and reference groups are estimated concurrently, the IRT LR test compares the likelihood when a studied item's parameters are constrained to be invariant across groups with the likelihood function when the parameters of the studied item are allowed to vary across groups. The LR can be expressed as

$$G^2 = -2 \ln\left[\frac{L(A)}{L(C)}\right] \tag{3.11}$$

where *L*(A) is the likelihood obtained when the studied item's parameter(s) are freely estimated across groups (i.e., augmented model) and *L*(C) is the likelihood obtained when the studied item has parameter(s) constrained to equality across groups (i.e., constrained model). The G^2 statistic is distributed approximately as a chi-square statistic with degrees of freedom equal to the number of constraints needed to derive the constrained model from the augmented model. The null hypothesis is that the item parameters are invariant

across groups. The LR test can be applied to both dichotomous and polytomous data (Millsap & Everson, 1993; Thissen, 2001).

3.4 Procedure

The dissertation study was structured as follows:

1. Twelve hundred data sets were generated using Mplus 6.12 for each condition.
2. The first 1,000 data sets without data issues and complexities (e.g., response options not selected) were analyzed (Paxton et al., 2001).
3. IRTLRDIF v2.0b (Thissen, 2001) and the multiple group function in the R package MIRT (R Core Team, 2012) were used to analyze data using the GR IRT model and a LR test of DIF for each item individually. The number of quadrature points used during estimation were 49 (P. Chalmers, personal communication, June 30, 2014) and 42 (D. Thissen, personal communication, July 4, 2014), respectively.
4. Mplus 6.12 was used to analyze data using a LCFA model and a chi-square nested model test of DIF for each item individually.
5. Mplus 6.12 was used to analyze data using a CCFA model and a chi-square nested model test of DIF for each item individually.
6. The output of each analysis was stored in a data file.
7. The process looped over all 1,000 simulated data sets.

8. All output data files were read into SAS® 9.3, and dependent variables were created and stored for the specific condition.
9. The process looped over all conditions of the study.
10. The results of the data simulation were interpreted and discussed.

3.5 Data Generating Model

The CCFA model was used to simulate data. Data were simulated using the Monte Carlo feature in Mplus 6.12. A criticism of simulation studies is that parameter values are chosen for mathematical convenience, and therefore may not reflect empirical data (Paxton et al., 2001). In an attempt to increase the external validity of the dissertation study, the population parameters were based on real data from an analysis done by Love (2012).

3.6 Data Analysis

Mplus 6.12, IRTLRDIF version 2.0b (Thissen, 2001), and the multiple group () function in the R package MIRT (R Core Team, 2012) were used to analyze the data. Data were analyzed using the GR IRT model coupled with a LR test for DIF (Thissen, 2001), and a LCFA and CCFA model both coupled with a comparable chi-square test of nested models. Customarily, factorial invariance is assessed by testing an entire parameter matrix, whereas DIF is assessed by testing each item individually (Millsap, 2011). In the current study, however, both factorial invariance and DIF were assessed by testing each item individually. This was done so that the results of the CFA and IRT

models' analyses could be directly compared**.** Using the LCFA model, each DIF test is a nested model comparison with two degrees of freedom. When using the CCFA model and the GR IRT model, each DIF test is a nested model comparison with five degrees of freedom. Because the IRT LR test statistic follows a chi-square distribution (Meade & Lautenschlager, 2004), this approach makes the assessment of measurement invariance across measurement frameworks identical (Stark et al., 2006). Finally, the mean and variance of the common factor of the reference group were fixed to 0 and 1, respectively, whereas, the mean and variance of the common factor of the focal group were freely estimated.

The dependent variables for the dissertation study were the TP and FP rates for each condition averaged across all replications. The TP rate was calculated by taking the number of items simulated to have DIF that were successfully detected as DIF items divided by the total number of DIF items generated. The FP rate was calculated by taking the number of items flagged as DIF items divided by the total number of items simulated to not contain DIF. Both the TP and FP rates were averaged over all replications within each condition.

3.7 Manipulated Factors

3.7.1 Source of DIF

Research suggests that both CFA and IRT have a difficult time identifying DIF associated with an item's factor loading (i.e., discrimination) parameter only (Kim &

Yoon, 2011; Meade & Lautenschlager, 2004). When examining the power of the methods to identify DIF associated with a polytomous item's threshold parameter, Meade and Lautenschlager (2004) found that IRT outperformed LCFA, whereas Kim and Yoon (2011) found that the CCFA and IRT performed similarly. A noteworthy point is that the studies may not be comparable because different CFA models were fit to the data and the source of DIF across the studies vary greatly. In an attempt to remedy this issue, this study included the source of DIF levels used by Kim and Yoon (2011), Meade and Lautenschlager (2004), and Stark et al. (2006). By including the sources of DIF used by all of the studies within a single study (see Table 3.2), the results of the current study may be compared to the aforementioned studies.

3.7.2 Size of DIF

Previous studies have simulated various amounts of DIF between the reference and focal groups (e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006; Woods, 2008, 2009). Woods (2008, 2009) operationalized DIF by creating a difference in item parameter estimates ranging from .3-.7. Stark et al. (2006) operationalized small and large DIF as an item factor loading difference of 0.15 and threshold difference of 0.25 and an item factor loading difference of 0.4 and threshold difference of 0.5, respectively (p. 1295). Meade and Lautenschlager (2004), however, simulated DIF using a constant of 0.4. Clearly, the values of small and large DIF simulated in previous research vary. In an attempt to add consistency across studies, the

values simulated by Kim and Yoon (2011) were chosen. Consequently, small DIF was operationalized as a factor loading difference of 0.2 and a threshold difference of 0.3, and large DIF was operationalized as a factor loading difference of 0.4 and a threshold difference of 0.6 (see Table 3.2). Both conditions favored the reference group.

3.7.3 Sample Size

Research suggests that the ability of the IRT model to detect items exhibiting DIF is related to the sample size of both groups (Woods, 2009b). As a consequence, this dissertation study manipulated sample size. Previous studies manipulated the sample size factor by including sample sizes of 150, 500, and 1,000 (Meade & Lautenschlager, 2004), 500 and 1,000 (Stark et al., 2006), and 200, 500, and 1,000 (Kim & Yoon, 2011). The current dissertation study studied the effect of sample size by including sample sizes of 250, 500, and 1,000.

3.7.4 Baseline Model

There are multiple ways to identify the IRT, LCFA, and CCFA models (Meade & Wright, 2012; Stark et al., 2006). Two common approaches to identifying the IRT, LCFA, and CCFA models are to use a free-baseline model or a constrained-baseline model. Among the three simulation studies that compared unobserved conditional invariance techniques to identifying DIF (i.e., Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006), only Stark et al. (2006) studied the effect of using a free-baseline model versus a constrained-baseline model on identifying an item

exhibiting DIF. For these reasons, research is needed that compares the IRT, LCFA, and CCFA models using both types of baseline models (i.e., free-baseline and constrained-baseline models).

This dissertation addressed this gap by including both free- and constrained-baseline modeling approaches. The free-baseline model approach identified the model by constraining a single referent item to be equal across groups and allowing all remaining items to be freely estimated across groups, and thus assumes that configural invariance holds and the referent item is invariant across groups (Millsap, 2011; Stark et al., 2006). Given the free-baseline model, each item was studied for DIF by comparing the free-baseline model to a compact model, which constrains the studied item to equality across groups. The constrained-baseline model approach, however, identified the model by constraining all items on the scale to equality across groups, and thus assumes that configural invariance holds and all items are invariant. Given the constrained-baseline model, each item was studied for DIF by comparing the constrained-baseline model to an augmented model, which freely estimated the studied item across groups. It is important to note that both the free- and constrained-baseline models fix the mean and variance of the latent factor of the reference group to zero and one, respectively, and freely estimates the mean and variance of the latent factor of the focal group.

As previously discussed, there are multiple baseline models that can be used to identify the IRT, LCFA, and CCFA models (Meade & Wright, 2012; Stark et al., 2006;

Woods, 2009a). These include the constrained-baseline model, free-baseline model, and a baseline model (i.e., anchor) that is iteratively purified (Meade & Wright, 2012; Woods, 2009a). The accuracy and ease of implementation and use of a baseline model are taken into consideration when deciding which baseline model to use (Meade & Wright, 2012). In the current study, only the constrained-baseline and free-baseline models are included as a manipulated factor. The constrained-baseline model approach is the default setting for IRTLRDIF v2.0b (Thissen, 2001) and MPLUS 6.12, and, as a consequence, is assumed to be used most often by applied researchers because it is easy to implement (Kim & Yoon, 2011). Likewise, the free-baseline model is included in the study because it is easy to program and non-iterative (Woods, 2009a). Purification methods, however, require multiple analyses, and thus are not easy to implement and use. In addition, it is unclear if the additional complexity of the purification methods lead to a significant increase in the accuracy of DIF detection when compared to a free-baseline model given a 5-item scale. For these reasons, baseline models based upon purification methods were not included in this study.

3.8 Dependent Variables

3.8.1 True Positive Rate

Previous studies compared the TP rates of CFA and IRT in DIF detection (e.g., Kim & Yoon, 2011). In this study, the TP rate was calculated as the proportion of DIF items correctly identified as having DIF across the number of replications in each

condition. The TP rates were averaged over all replications in each condition included in the study.

3.8.2 False Positive Rate

Previous studies compare the FP rates of CFA and IRT in DIF detection (e.g., Kim & Yoon, 2011). In this study, the FP rate represents the proportion of times an item having no DIF was incorrectly flagged as having DIF across the number of replications in each condition. The FP rates were averaged over all replications in each condition included in the study. Similar to Elosua (2011), Kim and Yoon (2011), and Stark et al. (2006), it is assumed that the nominal FP rate is .05.

3.9 Decision Rule

Each model will have a TP and FP rate within each condition of the study (i.e., source of DIF factor × size of DIF factor × sample size factor × baseline model factor). If the TP rate is at or above the nominal rate of .95, the model receives a "1" for the TP rate category; whereas if the TP rate is below the nominal rate of .95, the model receives a "0" for the TP rate category. If the FP rate is at or below the nominal rate of .05, the model receives a "1" for the FP rate category; whereas if the FP is above the nominal rate of .05, the model receives a "0" for the FP rate category. Thus, there are a total of four possible outcomes given the coding scheme. The following list is rank-ordered based on the assumption that falsely identifying an item as exhibiting DIF (i.e., FP > .05) is preferred over falsely identifying an item as being invariant (i.e., TP < .95).

The first and most preferred outcome is for a model to have a TP rate that is at or above .95 and a FP rate that is at or below .05. If more than one model meets this criteria within a given condition, the model with the FP rate at or closest to .05 is preferred. If more than one model has a FP rate at .05, the model within this subset of models with the largest TP rate is preferred.

The second outcome is preferred when none of the models meet the requirements of the first outcome. The second outcome is for a model to have a TP rate that is at or above .95 and a FP rate that is above .05. If more than one model meets this criteria within a given condition, the model within this subset of models with the smallest FP rate is preferred.

Assuming that no model meets the first and second outcomes, the third outcome is preferred. The third outcome is for a model to have a TP rate that is below .95 and a FP rate that is at or below .05. If more than one model meets this criteria within a given condition, the model within this subset of models with largest TP rate is preferred. If two or more models have the same TP rate, the model with within this subset of models the FP rate that is closest to .05 preferred.

The fourth and final outcome is preferred when no model meets the requirements of the first three outcomes. The fourth outcome is for a model to have a TP rate that is below .95 and a FP rate that is above .05. If more than one model meets this criteria within a given condition, both the TP and FP rates are considered. Generally speaking,

the model with the smallest FP rate is preferred. This assumes, however, that there is not much variation in the TP rates across models.

Once the preferred model is chosen for each condition (i.e., source of DIF factor × size of DIF factor × sample size factor × baseline model factor), the results are aggregated across sample sizes within each level of the size of DIF factor given a specific source of DIF and baseline modeling approach. The most precise model is the model with the highest frequency of being chosen as the preferred model at the aggregated level. In some instances, the same model is preferred across the various levels of the size of DIF factor. In these situations, the results are aggregated across the size of DIF factor and a single model is presented as being most precise given a specific baseline modeling approach.

CHAPTER FOUR

RESULTS

The results of the simulation study described in Chapter 3 are presented in this chapter. First, results are presented to demonstrate the validity of the generated data within each condition. Second, the results from each condition are presented. Within each condition, the results under small DIF are presented, followed by the results under large DIF. Within each level of the size of DIF factor, the FP rates are first presented, then TP rates are presented. Given the dependent variable of interest, the impact of sample size and baseline model are discussed. Prior to the discussion of the results of the simulation study, the details of the simulation study are briefly discussed.

4.1 Details of the simulation study

Similar to most simulation studies, the data generation and analysis procedures did not go exactly as planned (Paxton et al., 2001). To be specific, during the data generation phase, some of the simulated data sets had issues that warranted removal of the data set. As mentioned by Paxton et al. (2001), it is likely that some of the data sets need to be replaced by chance. Given the categorical nature of the data and sample sizes of 250 per group, there were some instances where data sets were created with response options unselected. Given the assumptions of the GR model (Samejima, 1969), this causes many issues pertaining to data analysis. The GR model assumes that a higher level of ability is needed as the response options progress along the ability distribution. By

having an empty cell, the data suggest that the response option is not needed. Instead of collapsing the response option, additional data sets were created, as suggested by Paxton et al., and used to replace the data sets with empty response option cells.

After removing all data sets with empty cells, there was still an additional issue that remained. Some of the data sets had issues where the model did not converge (e.g., negative variance). As previously discussed, even though only 1,000 data sets were needed for the analysis, 1,200 data sets were created (Paxton et al., 2001). Whenever an issue arose during data analysis, similar to the data generation phase, the data set was replaced.

4.2 Results of Simulation Study

4.2.1 Condition 1

The data for condition 1 were simulated so that the parameters were the same across groups (i.e., no DIF). Thus, only the FP rates were of interest. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.1. The 95% coverage rate is the proportion of replications that have a 95% confidence interval that includes the population parameter (Muthén & Muthén, 1998-2010). Generally speaking, the parameters were within the generating model parameters up to the second decimal place (i.e., 0.00). In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.1

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 1

Description	*N*	Average Parameter Bias	Average 95% Coverage Rates
0 DIF Items	250	0.00	0.95
No λ DIF	500	0.00	0.95
No υ_c DIF	1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*.

Table 4.2 presents the results of condition 1. The FP rates remained constant at the nominal rate of .05 for the IRT model across all levels of the sample size factor regardless of the baseline model used. The FP rates of the CCFA model were similar across all levels of the sample size and baseline model factors. Likewise, the FP rates of the LCFA model exhibited little variation across the levels of the sample size and baseline model factors. Based on the results presented in Table 4.2, it is argued that the optimal model is the IRT model using either a constrained- or free-baseline model.

4.2.2 Condition 2

The data for condition 2 were simulated such that the factor loading parameters were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the relationship between the studied item and the latent construct of interest is higher for the reference group than it is for the focal group. Under condition 2, both the TP and FP rates

Table 4.2

False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 1

		Constrained-Baseline			Free-Baseline		
		IRT	LCFA	CCFA	IRT	LCFA	CCFA
Description	*N*	FP	FP	FP	FP	FP	FP
0 DIF Items	250	0.05	0.14	0.06	0.05	0.13	0.05
No λ DIF	500	0.05	0.13	0.06	0.05	0.12	0.06
No υ_c DIF	1000	0.05	0.12	0.05	0.05	0.12	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. FP stands for false positive rates. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*.

were of interest. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.3. Except for the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place (.00). It is important to note that this was a very small difference that may be due to noise. In addition, the average 95% coverage rates were all at the nominal rate of 95%.

Table 4.3

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 2

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
λ DIF		500	0.00	0.95
No υ_c DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .2. Large DIF equals .4.

The results of condition 2 are shown in Table 4.4. Beginning with the conditions where the degree of DIF was small, the FP rates of the IRT, LCFA, and CCFA models were above .05 when using a constrained-baseline model and, in general, increased as the sample size increased. The FP rates of the IRT, LCFA, and CCFA models were at or above .05 when using a free-baseline model and did not exhibit much variation across the

Table 4.4

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 2

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.63	0.06	0.66	0.16	0.70	0.10	0.58	0.05	0.57	0.14	0.58	0.05
λ DIF		500	0.95	0.06	0.90	0.18	0.96	0.13	0.91	0.06	0.84	0.14	0.89	0.05
No υ_c DIF		1000	1.00	0.08	1.00	0.20	1.00	0.22	1.00	0.05	0.99	0.13	0.99	0.06
	Large	250	0.99	0.07	0.99	0.18	1.00	0.17	0.98	0.06	0.97	0.14	0.99	0.06
		500	1.00	0.10	1.00	0.23	1.00	0.28	1.00	0.05	1.00	0.14	1.00	0.06
		1000	1.00	0.15	1.00	0.31	1.00	0.49	1.00	0.05	1.00	0.13	1.00	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .2. Large DIF equals .4.

various sample sizes. Finally, using a constrained-baseline model versus a free-baseline model led to higher FP rates for the IRT, LCFA, and CCFA models.

The TP rates of the models are also shown in Table 4.4. Given a sample size of 250, none of the models was able to correctly detect DIF at an adequate rate. As the sample size increased, the ability to detect DIF improved for all models. Nevertheless, only the IRT and CCFA models using a constrained-baseline were able to correctly detect DIF at an adequate rate given a sample size of 500. Using a constrained-baseline model versus a free-baseline model led to higher TP rates for the IRT, LCFA, and CCFA models. When the sample size reached 1,000, the ability to correctly detect DIF was similar for all models regardless of baseline-model.

Turning attention to the conditions where the data were simulated to exhibit large DIF, except for the IRT model using a free-baseline strategy and sample sizes of 500 and 1000, the FP rates were above the nominal rate of .05 for all of the studied models across all sample sizes (see Table 4.4). Given a constrained-baseline model, the FP rates of the IRT, LCFA, and CCFA models consistently increased as sample size increased. In contrast, there were only small differences, if any, in the FP rates of the IRT, LCFA and CCFA models using a free-baseline model across sample sizes.

Under large DIF, the performance of the studied models was very similar (see Table 4.4). The TP rates were at least .97 across all sample sizes and models. Taking all

results into consideration, it is argued that a free-baseline model is most accurate, and the IRT model is the most precise model using either baseline strategy.

4.2.3 Condition 3

The data for condition 3 were simulated such that all four of the latent threshold parameters (υ_1-υ_4) are different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this amounts to the focal group needing a higher amount of the latent construct of interest to select any of the five response options. As was the case with condition 2, the FP and TP rates were both of interest. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.5. Other than the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place (.00). In addition, the average 95% coverage rate was at the nominal rate of 95%.

The results of condition 3 are shown in Table 4.6. Under small DIF, the FP rates of the IRT, LCFA, and CCFA models using a constrained-baseline strategy were all well above the nominal rate of .05. The same holds true for the LCFA model given a free-baseline model. The FP rates of the IRT and CCFA models given a free-baseline approach, however, were at or slightly above .05. The FP rates of the free-baseline models were quite consistent across sample sizes; whereas the FP rates of the constrained-baseline models consistently increased as sample size increased. In addition,

using a constrained-baseline model led to higher FP rates than a free-baseline model for the IRT, LCFA, and CCFA models within each level of sample size.

Table 4.5

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 3

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
No λ DIF		500	0.00	0.95
υ_1-υ_4 DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

The TP rates of the models are also shown in Table 4.6. When the sample size is 250, only the LCFA model using a constrained-baseline approach correctly detected DIF at an adequate rate (TP = .96). Using a constrained-baseline model led to higher TP rates than a free-baseline model for the IRT, LCFA, and CCFA models. Once the sample size was 500 or larger, the TP rates were .99 and above for all of the studied models

Table 4.6

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 3

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.89	0.08	0.96	0.18	0.91	0.12	0.81	0.05	0.93	0.14	0.85	0.06
No λ DIF		500	1.00	0.12	1.00	0.22	1.00	0.17	0.99	0.05	1.00	0.14	0.99	0.05
υ_1-υ_4 DIF		1000	1.00	0.21	1.00	0.31	1.00	0.30	1.00	0.05	1.00	0.13	1.00	0.05
	Large	250	1.00	0.21	1.00	0.25	1.00	0.31	1.00	0.05	1.00	0.13	1.00	0.05
		500	1.00	0.41	1.00	0.38	1.00	0.58	1.00	0.05	1.00	0.13	1.00	0.06
		1000	1.00	0.73	1.00	0.62	1.00	0.87	1.00	0.05	1.00	0.14	1.00	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

regardless of baseline strategy.

Under large DIF, only the IRT and CCFA models using a free-baseline approach falsely identified an item as having DIF at an adequate rate (see Table 4.6). The FP rates of the IRT, LCFA, and CCFA models using a constrained-baseline strategy consistently increased as the sample size increased. The FP rates of the IRT, LCFA, and CCFA models using a free-baseline approach, however, did not exhibit a consistent change across sample sizes. The results also illustrate that using a constrained-baseline strategy versus a free-baseline strategy led to an increase of the FP rate within each sample size for all of the studied models.

Under large DIF, the TP rates reached 1.00 for all models across all levels of sample size regardless of baseline strategy (see Table 4.6). Thus, a comparison of the TP rates of the studied models across sample size (i.e., 250, 500, and 1,000) and baseline model (i.e., free-baseline and constrained-baseline) suggests that they all have comparable power to accurately detect an item exhibiting DIF. Taking all results into consideration, it is concluded that the IRT model is most precise regardless of the baseline approach. Considering the similar FP and TP rates of the IRT and CCFA models using a free-baseline strategy, both models are effective at DIF detection.

4.2.4 Condition 4

The data for condition 4 were simulated such that the fourth (υ_4) of four latent threshold parameters (υ_1-υ_4) was different across groups. Assuming a Likert-scaled item

with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this amounts to the focal group needing a higher amount of the latent construct of interest to select a response of Strongly Agree. Finally, the FP and TP rates were the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.7. Other than the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.7

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 4

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
No λ DIF		500	0.00	0.95
υ_4 DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

The results of condition 4 are shown in Table 4.8. Beginning with the conditions where the degree of DIF is small, the FP rates were consistently above the nominal rate of .05 for the IRT and CCFA models using a constrained-baseline approach and the LCFA model using both baseline model approaches. The FP rates of the IRT and CCFA models using a free-baseline strategy were similar and, at times, at the nominal level of .05. For all of the models using a constrained-baseline model, as the sample size increased, the FP rate also increased. Given the free-baseline models, there was not a consistent pattern of FP rates across sample size. Finally, the data suggest that using a constrained-baseline model versus a free-baseline model led to an increase in FP rates.

The TP rates of the models are also shown in Table 4.8. The IRT model outperformed the LCFA and CCFA models at detecting DIF regardless of the baseline strategy. Generally speaking, an increase in sample size led to an increase in TP rates. Focusing on the IRT and LCFA models, using a constrained-baseline approach versus a free-baseline approach resulted in a slight increase in TP rates. Given the CCFA model, there was little difference in TP rates between the constrained-baseline strategy and free-baseline strategy.

When the data were simulated to exhibit large DIF, the FP rates of the IRT and CCFA models using a free-baseline approach FP were consistent (see Table 4.8). The IRT-C, LCFA-C, CCFA-C, and LCFA-F models' FP rates ranged from .10-.59. As the sample size increased, the FP rates also increased for the IRT, LCFA, and CCFA models

Table 4.8

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 4

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.78	0.07	0.42	0.15	0.58	0.06	0.69	0.05	0.37	0.14	0.59	0.06
No λ DIF		500	0.99	0.09	0.71	0.16	0.92	0.07	0.96	0.05	0.64	0.13	0.92	0.05
υ_4 DIF		1000	1.00	0.15	0.95	0.17	1.00	0.09	1.00	0.04	0.90	0.13	1.00	0.06
	Large	250	1.00	0.17	0.97	0.18	1.00	0.10	1.00	0.05	0.93	0.14	0.99	0.06
		500	1.00	0.31	1.00	0.21	1.00	0.12	1.00	0.06	1.00	0.13	1.00	0.06
		1000	1.00	0.59	1.00	0.29	1.00	0.21	1.00	0.05	1.00	0.12	1.00	0.05

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

using a constrained-baseline strategy. There was no consistent pattern in the FP rates of the IRT, LCFA, and CCFA models using a free-baseline strategy across sample sizes. Lastly, the constrained-baseline approach led to higher FP rates for each of the studied models than the free-baseline approach.

Under large DIF, the ability to correctly identify an item as exhibiting DIF was similar for the IRT, LCFA, and CCFA models across all sample sizes and baseline strategies (see Table 4.8). Given the homogenous results presented in Table 4.8, it is difficult to conclude that there were any differences in TP rates related to the sample size and/or baseline model for the IRT and CCFA models. Perhaps an argument could be made for the LCFA model given a sample of 250. The results in Table 4.8 suggest that the IRT model is the most precise model using a free-baseline strategy. The most precise model, which is either the IRT model or CCFA model, using a constrained-baseline strategy depends upon the size of DIF, however.

4.2.5 Condition 5

Data for condition 5 were simulated such that the last two (υ_3- υ_4) of four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the focal group needs a higher amount of the latent construct of interest to select a response of Agree or Strongly Agree. The FP and TP rates are the dependent variables. The average parameter bias and 95% coverage rates across

simulated data sets are presented in Table 4.9. Other than the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.9

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 5

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
No λ DIF		500	0.00	0.95
υ_3 and υ_4 DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

The results of condition 5 are shown in Table 4.10. First, the cells where the degree of DIF was small are discussed. The FP rates were above the nominal rate of .05 for all of the studied models using a constrained-baseline strategy, and the LCFA model using a free-baseline strategy. The FP rates for the IRT and CCFA models using a

Table 4.10

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 5

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.89	0.09	0.81	0.16	0.82	0.09	0.81	0.06	0.74	0.14	0.77	0.06
No λ DIF		500	0.99	0.12	0.98	0.18	0.99	0.12	0.98	0.05	0.96	0.13	0.97	0.06
υ_3 and υ_4 DIF		1000	1.00	0.18	1.00	0.22	1.00	0.16	1.00	0.05	1.00	0.14	1.00	0.05
	Large	250	1.00	0.22	1.00	0.23	1.00	0.18	1.00	0.06	1.00	0.14	1.00	0.07
		500	1.00	0.39	1.00	0.30	1.00	0.31	1.00	0.05	1.00	0.12	1.00	0.06
		1000	1.00	0.70	1.00	0.51	1.00	0.55	1.00	0.04	1.00	0.12	1.00	0.04

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

free-baseline strategy were at or slightly above the nominal rate of .05. For all of the models using a constrained-baseline, as the sample size increased, the FP rate also increased. Given a free-baseline strategy, there was a slight decrease in FP rates as sample size increased for the IRT and CCFA models. There was no consistent pattern in FP rates across sample sizes given the LFCA model. Lastly, using a constrained-baseline model versus a free-baseline model led to a slight increase in the FP rate of the IRT, LCFA, and CCFA models.

The TP rates of the models are also shown in Table 4.10. Given small DIF, the IRT, LCFA, and CCFA models were unable to correctly detect DIF at a sufficient rate until the sample size reached at least 500 for both baseline approaches. As sample size increased, so did the TP rates of all of the models for both baseline models. In addition, using a constrained-baseline strategy led to higher TP rates than using a free-baseline strategy.

Turning attention to the cells simulated to exhibit large DIF, even though none of the studied models maintained the nominal FP rate of .05 across sample sizes, the IRT model using a free-baseline strategy performed well given a sample size of 500 and above (see Table 4.10). Given the constrained-baseline models, as the sample size increased, the FP rates also increased. Conversely, given the IRT and CCFA models using a free-baseline approach, as the sample size increased, the FP rates slightly

decreased. There was not a consistent pattern given the LCFA model using a free-baseline approach.

Examining the TP rates of the cells simulated to exhibit large DIF, the TP rates were all 1.00 (see Table 4.10). This suggests that the IRT, LCFA, and CCFA models have the power to detect large DIF across all levels of the sample size factor regardless of baseline model. Based on the results presented in Table 4.10, the IRT model using a constrained-baseline was most accurate under conditions of small DIF, whereas the CCFA model using a constrained-baseline was most accurate under conditions of Large DIF. The IRT model, however, was clearly most precise using a free-baseline model.

4.2.6 Condition 6

Data for condition 6 were simulated such that the first and last (υ_1 and υ_4, respectively) of four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the focal group is less likely to select an extreme response of Strongly Disagree or Strongly Agree. The FP and TP rates are the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.11. Other than the data simulated with small DIF and a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.11

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 6

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
No λ DIF		500	0.00	0.95
υ_1 and υ_4 DIF		1000	0.00	0.95
	Large	250	0.00	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

The results of condition 6 are shown in Table 4.12. Under small DIF, the FP rate of the LCFA and CCFA models were above the nominal rate of .05 across all levels of sample size. The FP rate of the IRT model using a free-baseline model, however, only exceeded the nominal rate of .05 when the sample size is 250. Interestingly, among the constrained-baseline models, only the FP rates of the IRT model increased as sample size increased. There was not a consistent pattern of FP rates for the remaining models across sample sizes. Finally, using a constrained-model versus a free-baseline model led to an increase in FP rates.

Examining the TP rates of the studied models, it is apparent that the power of the

Table 4.12

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 6

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.85	0.09	0.41	0.16	0.68	0.08	0.78	0.06	0.36	0.14	0.68	0.07
No λ DIF		500	0.99	0.11	0.69	0.15	0.96	0.08	0.97	0.05	0.62	0.14	0.96	0.06
υ_1 and υ_4 DIF		1000	1.00	0.18	0.94	0.17	1.00	0.09	1.00	0.05	0.90	0.13	1.00	0.06
	Large	250	1.00	0.18	0.96	0.17	1.00	0.08	1.00	0.05	0.93	0.13	1.00	0.05
		500	1.00	0.34	1.00	0.21	1.00	0.12	1.00	0.06	1.00	0.13	1.00	0.06
		1000	1.00	0.64	1.00	0.29	1.00	0.18	1.00	0.04	1.00	0.14	1.00	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals .3. Large DIF equals .6.

models varied greatly (see Table 4.12). Given a sample size of 250, none of the models' TP rate exceeded a rate of .85. The performance of the LCFA models were poorer than those of the IRT and CCFA models across all sample sizes. The data suggests that an increase in sample size led to an increase in TP rates. Focusing on the IRT and CCFA models, using a constrained-baseline strategy led to an increase of TP rates. Interestingly, given the CCFA model, there was no difference in TP rates using the constrained-baseline model versus the free-baseline model.

Turning attention to large DIF, the FP rates of the constrained-baseline models and the LCFA model using a free-baseline strategy exceeded the nominal value of .05 across all levels of the sample size factor (see Table 4.12). As the sample size increased, FP rates of the constrained-baseline models also increased and grew in variation. The FP rates of the free-baseline model, however, were similar across all sample sizes. Using a constrained-baseline model versus a free-baseline model led to an increase in FP rates within each model.

Examining the TP rates of the studied models under large DIF, it is important to note that the TP rates of the IRT, LCFA, and CCFA models were similar across all levels of sample size and baseline strategy (see Table 4.12). Because the TP rates of the studied models did not exhibit consistent variation, the impact of using a larger sample or constrained- versus free- baseline model could not be determined. In conclusion, the IRT model slightly outperformed the CCFA model using a free-baseline strategy. Given a constrained-baseline strategy, it is argued that, in general, the CCFA model is most accurate.

4.2.7 Condition 7

Data for condition 7 were simulated such that the factor loading parameter (λ) and all four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the relationship between the item and the latent construct is stronger for the reference group than the focal group (i.e., λ DIF) and the focal group needs a higher amount of the latent construct of interest to select any of the five response options. The FP and TP rates were the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.13. Other than the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95% for all conditions.

The results of condition 7 are shown in Table 4.14. First, the cells where the degree of DIF is small are discussed. The FP rate of the IRT model using a free-baseline approach remained at .05 across all sample sizes, whereas the CCFA model using a free-baseline approach had minor fluctuations around the nominal value of .05 across sample sizes. The remaining models' FP rates were consistently above the nominal rate of .05. This suggests that using a constrained-baseline versus a free-baseline model led to higher FP rates. The FP rates of the constrained-baseline models increased as sample size increased. Given the free-baseline models, there was little variation, if any, in the FP rates as the sample size increased.

Examining the TP rates of the studied models, it is apparent that the ability to correctly detect DIF of the models were very similar (see Table 4.14). Given a sample

size of 250, the constrained-baseline models outperformed the free-baseline models. Of the free-baseline models, only the CCFA model detected DIF at an acceptable rate. When the sample size was at least 500, the TP rates were 1.00 for all of the models. As a consequence, it is difficult to determine the impact that sample size and baseline model had on these models under these conditions.

Table 4.13

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 7

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
λ DIF		500	0.00	0.95
υ_1-υ_4 DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

Next, the FP rates under large DIF are discussed. Even though the performance of the IRT and CCFA models using a free-baseline approach were similar, only the IRT-F model had a FP rate that did not exceed .05 across all sample sizes (see Table 4.14). The FP rates of the constrained-baseline models and LCFA model using a free-baseline strategy were above the nominal rate of .05 across all sample sizes. The FP rates of the IRT, LCFA, and CCFA models increased as sample size increased given a

Table 4.14

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 7

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.97	0.09	0.96	0.17	0.99	0.15	0.92	0.05	0.93	0.13	0.96	0.06
λ DIF		500	1.00	0.13	1.00	0.17	1.00	0.27	1.00	0.05	1.00	0.13	1.00	0.05
υ_1-υ_4 DIF		1000	1.00	0.23	1.00	0.22	1.00	0.45	1.00	0.05	1.00	0.14	1.00	0.06
	Large	250	1.00	0.19	1.00	0.18	1.00	0.43	1.00	0.05	1.00	0.13	1.00	0.05
		500	1.00	0.36	1.00	0.24	1.00	0.71	1.00	0.05	1.00	0.13	1.00	0.06
		1000	1.00	0.67	1.00	0.37	1.00	0.94	1.00	0.05	1.00	0.14	1.00	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6

constrained-baseline model. On the contrary, given a free-baseline model, there was not a consistent pattern of FP rates across sample sizes. In addition, the FP rates were much higher for the constrained-baseline models than the free-baseline models across all levels of the sample size factor.

Focusing on the TP rates of the IRT, LCFA, and CCFA models under large DIF, all of the models have a TP rate at 1.00 across all sample sizes regardless of baseline model (see Table 4.14). As a consequence, the impact of sample size and baseline model could not be determined.

Even though the performance of the IRT and CCFA models using a free-baseline are almost identical, generally speaking, the data suggest that the IRT model is most accurate. Using a constrained-baseline model, however, the conclusion depends on the size of DIF. Under small DIF, the IRT model was most accurate, whereas the LCFA model was most accurate under large DIF. On average, the LCFA is preferred. Clearly, the free-baseline strategy is more precise than the constrained-baseline strategy.

4.2.8 Condition 8

Data for condition 8 are simulated such that the factor loading parameter (λ) and the fourth (υ_4) of four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the relationship between the item and the latent construct is stronger for the reference group than the focal group (i.e., λ DIF) and the focal group needs a higher amount of the latent construct of interest to the Strongly Agree response option. The FP and TP rates were the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented

in Table 4.15. Other than the data simulated with a sample size of 250 (-.01), the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.15

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 8

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
λ DIF		500	0.00	0.95
υ_4 DIF		1000	0.00	0.95
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

The results of condition 8 are shown in Table 4.16. Under small DIF, the FP rates of the constrained-baseline models were quite similar and above the nominal rate of .05. Even though the FP rates of the IRT and CCFA models using a free-baseline model models were also similar, only the IRT model's FP rate remained at .05 across all sample sizes. As sample size increased, the FP rates of the IRT, LCFA, and CCFA models using a constrained-baseline strategy also increased. The FP rates of the free-baseline models exhibited little variation, if any, across all sample sizes. Based on the results presented in Table 4.16, using a constrained-model compared to a free-baseline model led to an

Table 4.16

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 8

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.90	0.10	0.87	0.17	0.94	0.12	0.77	0.05	0.83	0.13	0.88	0.05
λ DIF		500	1.00	0.14	1.00	0.19	1.00	0.16	0.98	0.05	0.99	0.13	1.00	0.06
υ_4 DIF		1000	1.00	0.25	1.00	0.25	1.00	0.27	1.00	0.05	1.00	0.13	1.00	0.06
	Large	250	1.00	0.21	1.00	0.20	1.00	0.22	1.00	0.06	1.00	0.14	1.00	0.06
		500	1.00	0.39	1.00	0.27	1.00	0.37	1.00	0.05	1.00	0.13	1.00	0.05
		1000	1.00	0.71	1.00	0.41	1.00	0.64	1.00	0.04	1.00	0.13	1.00	0.05

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

inflation of FP rates.

Examining the TP rates of the studied models, it is apparent that the power of the models are very similar (see Table 4.16). A sample size of 500 was needed for adequate DIF detection for the IRT, LCFA, and CCFA models regardless of baseline model. Increasing the sample size from 250 to 500 clearly led to an increase of TP rates for all models. Because there is little variation in TP rates after increasing the sample size from 500 to 1,000, it is assumed that the improved performance due to sample size is minor once the sample size is about 500. The same assumption is made pertaining to the use of a constrained- versus free-baseline model.

Given large DIF, the FP rates of the constrained-baseline model ranged from .20-.71 (see Table 4.16). The FP rates of the free-baseline models, however, did not exceed .14. This suggests that a constrained-baseline model deteriorates the performance of the IRT, LCFA, and CCFA models as compared to using a free-baseline approach. Additionally, an increase of sample size led to an increase in the FP rates of the constrained-baseline models. There was not much of a difference, if any, in FP rates amongst free-baseline models as sample size increased; the FP rates of the IRT, LCFA and CCFA models either stay the same or decrease.

Given large DIF, the TP rates of the IRT, LCFA, and CCFA models were all 1.00 regardless of baseline model strategy (see Table 4.16). Thus, it was impossible to detect the impact of increasing the sample size and using a constrained- versus free-baseline model under large DIF. This does suggest that any of the models can be employed given large DIF.

In summary, given a free-baseline model, either the IRT or CCFA model will suffice. It is difficult to identify the most effective model using a constrained-baseline model, as it depends on the size of DIF. In the presence of small DIF, the most precise model varies across sample sizes. Under the condition of large DIF, the LCFA model is the preferred constrained-baseline model.

4.2.9 Condition 9

Data for condition 9 were simulated such that the factor loading parameter (λ) and the third (υ_3) and fourth (υ_4) of four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the relationship between the item and the latent construct is stronger for the reference group than the focal group (i.e., λ DIF) and the focal group needs a higher amount of the latent construct of interest to select the Agree and Strongly Agree response options. The FP and TP rates were the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.17. Other than the data simulated with a sample size of 250 (-.01) and small DIF with a sample size of 500, the parameters were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

The results of condition 9 are shown in Table 4.18. Given small DIF, the FP rates of the IRT model using a free-baseline model were at the nominal value of .05 across all levels of sample size, whereas the FP rate of the CCFA model using a free-baseline model did not reach the nominal value of .05 until the sample size was 1,000. The FP rates of the constrained-baseline models and the LCFA model using a free-baseline

model, however, were above the nominal value of .05 across all levels of the sample size factor. Given the constrained-baseline models, the FP rates increased as sample size increased. Given the free-baseline models, there was not enough variation in FP rates to attribute a change associated with increasing sample size. Lastly, using a constrained-baseline model versus a free-baseline model led to an increase in FP rates.

Table 4.17

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 9

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
λ DIF		500	-0.01	0.95
υ_3 and υ_4		1000	0.00	0.95
DIF				
	Large	250	-0.01	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

The TP rates of the studied models under small DIF are shown in Table 4.18. Given a sample size of 250, TP rates of the constrained baseline models were higher than those of the free-baseline models. This suggests that using a constrained-baseline model versus a free-baseline model led to slightly higher TP rates in small samples (i.e., 250). When the sample size reached 500, the TP rates of all the studied models were 1.00

Table 4.18

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 9

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.96	0.10	0.95	0.16	0.97	0.13	0.89	0.05	0.92	0.14	0.94	0.06
λ DIF		500	1.00	0.15	1.00	0.19	1.00	0.20	1.00	0.05	1.00	0.14	1.00	0.06
υ_3 and υ_4 DIF		1000	1.00	0.26	1.00	0.25	1.00	0.34	1.00	0.05	1.00	0.13	1.00	0.05
	Large	250	1.00	0.22	1.00	0.21	1.00	0.30	1.00	0.05	1.00	0.14	1.00	0.05
		500	1.00	0.42	1.00	0.28	1.00	0.52	1.00	0.05	1.00	0.15	1.00	0.06
		1000	1.00	0.72	1.00	0.44	1.00	0.80	1.00	0.05	1.00	0.13	1.00	0.06

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

suggesting that an increase of sample size led to an increase in TP rates.

Given large DIF, only the IRT and CCFA models using a free-baseline model had FP rates that were at or near the nominal value of .05 across all sample sizes (see Table 4.18). Increasing the sample size led to an increase in the FP rates of the IRT, LCFA, and CCFA models using a constrained-baseline strategy. As a matter of fact, given a sample size of 1,000, the constrained-baseline models FP rates are .44 and above. On the contrary, the free-baseline models exhibited little variation and did not exceed .15. Thus, the effect of using a constrained-baseline model versus a free-baseline model was an increase of FP rates.

Given large DIF, the TP rates of all of the studied models are 1.00 (see Table 4.18). Because the TP rates are all the same, the effect of increasing the sample size could not be determined. In addition, the effect of using a constrained-baseline model versus a free-baseline model on TP rates could not be determined.

In conclusion, it is argued that the IRT model using a free-baseline model is most precise at DIF detection. Assuming that a constrained-baseline model is preferred, the most accurate model depends on the size of DIF. Given small DIF and a constrained-baseline model, the IRT model is preferred. Conversely, given large DIF, the LCFA model is preferred.

4.2.10 Condition 10

Data for condition 10 were simulated such that the factor loading parameter (λ) and the first (υ_1) and fourth (υ_4) of four latent threshold parameters (υ_1-υ_4) were different across groups. Assuming a Likert-scaled item with five response options (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), this occurs when the relationship

between the item and latent construct is stronger for the reference group than the focal group (i.e., λ DIF) and when the focal group is less likely to select an extreme response of Strongly Disagree or Strongly Agree. The FP and TP rates were the dependent variables. The average parameter bias and 95% coverage rates across simulated data sets are presented in Table 4.19. Other than the data simulated under the condition of small DIF and a sample size of 250 (-.01), the values were within the generating model parameters up to the second decimal place. In addition, the average 95% coverage rate was at the nominal rate of 95%.

Table 4.19

Parameter Bias and 95% Confidence Intervals for Data Simulated Under Condition 10

Description	Degree of DIF	*N*	Average Parameter Bias	Average 95% Coverage Rates
1 DIF Item	Small	250	-0.01	0.95
λ DIF		500	0.00	0.95
υ_1 and υ_4 DIF		1000	0.00	0.95
	Large	250	0.00	0.95
		500	0.00	0.95
		1000	0.00	0.95

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

The results of condition 10 are shown in Table 4.20. Even though the FP rates of the IRT and CCFA models using a free-baseline strategy were similar, only the IRT

Table 4.20

True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Under Condition 10

			Constrained-Baseline						Free-Baseline					
			IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Description	Degree of DIF	*N*	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1 DIF Item	Small	250	0.91	0.10	0.90	0.16	0.94	0.10	0.80	0.05	0.86	0.13	0.90	0.06
λ DIF		500	1.00	0.14	1.00	0.19	1.00	0.15	0.98	0.05	0.99	0.13	1.00	0.05
υ_1 and υ_4 DIF		1000	1.00	0.28	1.00	0.25	1.00	0.24	1.00	0.05	1.00	0.13	1.00	0.06
	Large	250	1.00	0.22	1.00	0.21	1.00	0.23	1.00	0.05	1.00	0.13	1.00	0.05
		500	1.00	0.43	1.00	0.30	1.00	0.38	1.00	0.05	1.00	0.14	1.00	0.06
		1000	1.00	0.73	1.00	0.44	1.00	0.62	1.00	0.05	1.00	0.12	1.00	0.05

Note. DIF stands for Differential Item Functioning. *N* stands for sample size. IRT stands for the Graded Response Item Response Theory model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. TP stands for true positive. FP stands for false positive. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. Small DIF equals λ = .2, υ = .3. Large DIF equals λ = .4, υ = .6.

model had FP rates at the nominal rate of .05 across all sample sizes. The FP rates given a constrained-baseline model, however, were all above the nominal value of .05. The same holds true for the LCFA model regardless of baseline model. Given the constrained-baseline models, the FP rate increased as sample size increased. In contrast, the FP rates of the IRT and LCFA models using a free-baseline model remained constant across all sample sizes. As illustrated in Table 4.20, using a constrained-baseline model versus a free-baseline model led to an increase in FP rates.

The TP rates under small DIF are presented in Table 4.20. Given a sample size of 250, none of the studied models' TP rates exceeded .94. When the sample size reached at least 500, however, the TP rates of all of the studied models exceeded .98. Given a sample size of 1,000, all of the models were able to correctly identify an item as exhibiting DIF at a rate of 1.00. The results also suggest that using constrained-baseline models versus a free-baseline model led to a slight increase in TP rates.

Given large DIF, only the IRT model using a free-baseline approach had FP rates that were at the nominal value of .05 (see Table 4.20). The FP rates of the CCFA model using a free-baseline approach showed slight variation around the nominal level of .05 at specific levels of sample size. Whereas the FP rates of the IRT, LCFA, and CCFA models using a constrained-baseline strategy and the LCFA model using a free-baseline were consistently larger than the nominal value of .05. Given the constrained-baseline models, increasing the sample size led to a consistent increase in FP rates. Given the IRT model using a free-baseline model, increasing the sample size did not lead to a change in FP rates. Because there was not a consistent pattern of FP rates across sample sizes for the LCFA and CCFA models using a free-baseline model, the impact of increasing the

sample size could not be determined. Lastly, the results suggest that using a constrained-baseline model versus a free-baseline model led to an increase of FP rates.

Given large DIF, the TP rates of all of the studied models were fixed at 1.00. Because the TP rates are all the same, the impact of increasing the sample size could not be determined. In addition, the impact of using a constrained-baseline model versus a free-baseline model on TP rates could not be determined. Taking all of the results into consideration, it is concluded that the CCFA model is the most precise constrained-baseline model given small DIF, and the LCFA model is the most precise constrained-baseline model given large DIF. Given a free-baseline model, the IRT and CCFA model are both precise. It is important to note, however, that the CCFA model is preferred given a sample size of 250.

4.3 Concluding Remarks

The goal of this section is to succinctly summarize the results. A reader that is interested in a detailed discussion and illustration of the results is referred to the sections above (i.e., Sections 4.2.1-4.2.10). In conclusion, the IRT model using a free-baseline strategy was the most accurate model across most of the 10 conditions regardless of sample size and size of DIF (see Table 4.21). It is important to note, however, that even though only conditions 8 and 10 have results that would provide an argument for the CCFA model using a free-baseline strategy being the most precise model (see Tables 4.16 and 4.20, respectively), the absolute differences in TP and FP rates between the IRT and CCFA model are small (i.e., .11 and .02, respectively) across all conditions (see Sections 4.2.1-4.2.10). Thus, in most instances a CCFA model will also suffice given a free-baseline approach.

Table 4.21

Average True Positive and False Positive Rates of IRT, LCFA, and CCFA Models Across Size of DIF and Sample Size Factors

		Constrained-Baseline						Free-Baseline					
		IRT		LCFA		CCFA		IRT		LCFA		CCFA	
Condition	Description	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
1	No DIF	**-**	**0.05**	-	0.13	-	0.06	**-**	**0.05**	-	0.12	-	0.06
2	λ and No υ_c DIF	**0.93**	**0.09**	0.92	0.21	0.94	0.23	**0.91**	**0.05**	0.90	0.14	0.91	0.06
3	No λ DIF and υ_1-υ_4 DIF	**0.98**	**0.29**	0.99	0.33	0.98	0.39	**0.97**	**0.05**	0.99	0.13	0.97	0.06
4	No λ DIF and υ_4 DIF	0.96	0.23	0.84	0.19	**0.92**	**0.11**	**0.94**	**0.05**	0.81	0.13	0.92	0.06
5	No λ DIF and υ_3-υ_4 DIF	0.98	0.28	0.96	0.27	**0.97**	**0.23**	**0.97**	**0.05**	0.95	0.13	0.96	0.06
6	No λ DIF and υ_1 and υ_4 DIF	0.97	0.26	0.83	0.19	**0.94**	**0.10**	**0.96**	**0.05**	0.80	0.13	0.94	0.06
7	λ and υ_1-υ_4 DIF	1.00	0.28	**0.99**	**0.23**	1.00	0.49	**0.99**	**0.05**	0.99	0.13	0.99	0.06
8	λ and υ_4 DIF	0.98	0.30	**0.98**	**0.25**	0.99	0.29	0.96	0.05	0.97	0.13	**0.98**	**0.05**
9	λ and υ_3- υ_4 DIF	0.99	0.31	**0.99**	**0.26**	1.00	0.38	**0.98**	**0.05**	0.99	0.14	0.99	0.06
10	λ and υ_1, and υ_4 DIF	0.98	0.31	**0.98**	**0.26**	0.99	0.29	0.96	0.05	0.97	0.13	**0.98**	**0.05**

Note. DIF stands for Differential Item Functioning. IRT stands for the Graded Response Item Response Theory Model. LCFA stands for the linear Confirmatory Factor Analysis model. CCFA stands for the categorical Confirmatory Factor Analysis model. λ represents the factor loading parameter. υ_c represents the threshold parameter for response option *c*. The most precise model given the condition and baseline model is shown in boldface.

Unfortunately, given a constrained-baseline model, the most precise model depends upon the source of DIF, size of DIF, and sample size. In general, the IRT model is most precise given small DIF and a constrained-baseline approach. As is the case using a free-baseline model, the results suggest that the CCFA model is most precise given condition 10 and small DIF. In contrast, given large DIF and a constrained-baseline approach, the LCFA model is most accurate across most of the conditions (see Table 4.21). This is due to the very large FP rates of the IRT and CCFA models (.07-.73 and .08-.94, respectively; see sections 4.2.1-4.2.10). Table 4.21 presents the results within each condition after averaging across the size of DIF and sample size factors.

CHAPTER FIVE

CONCLUSION

As discussed in Chapters 1 and 2, CFA and IRT are two unobserved conditional invariance approaches to assessing measurement invariance (Millsap & Everson, 1993). Even though both methods are designed to answer the same question, research suggests that that the two methods may lead to contrasting results (Meade & Lautenschlager, 2004). The inconsistent results were first brought to light by researchers comparing the CFA and IRT approaches using empirical data (e.g., Oishi, 2006; Raju et al., 2002; Reise et al., 1993). These inconsistencies led to a need for research comparing the CFA and IRT approaches using simulation studies. Kim and Yoon (2011), Meade and Lautenschlager (2004), and Stark et al. (2006) compared the CFA and IRT approaches to assessing measurement invariance using simulated data, and, ironically, added to the complexity of the issue. This was mostly due in part to the inconsistent design used across the three studies. Because of the inconsistent design across studies, additional research was needed that focuses on the existing gaps in the literature by comparing the LCFA, CCFA, and GR IRT models using 5-point Likert scaled data and a research design that encompasses the design of the aforementioned studies.

Chapter 3 presented the research questions and methods used to answer these research questions. Chapter 4 presented the results of the methods used to answer the research questions of interest. Chapter 5 presents the conclusion of the dissertation study. The remainder of Chapter 5 will present the following: (a) summary of the results, (b)

discussion of the results, (c) significance of the study, (d) limitations of the study, (e) recommendations for future work, and (f) final thoughts.

5.1 Summary of the Results

The dissertation research examined the performance of three unobserved conditional invariance techniques (i.e., LCFA, CCFA, and IRT) under 10 conditions (see Chapters 3 and 4 for a detailed discussion of conditions). This section briefly summarizes the results across all of the manipulated conditions presented in Chapter 4.

Based on the results of the dissertation study, it can be concluded that *the GR IRT model with a free-baseline is most precise when studying DIF using ordinal data across most conditions* (Stark et al., 2006). Given a constrained-baseline strategy, the most precise model depends on the source and size of DIF. Considering that an applied researcher does not know whether DIF is present and, if so, which item parameter is exhibiting DIF, it is recommended that applied researchers avoid using a constrained-baseline strategy.

This study also found that using a constrained-baseline model will inflate FP rates, and, as a consequence, lead to items being falsely identified as exhibiting DIF at a rate higher than expected by chance (Stark et al., 2006). The constrained-baseline model will also correctly identify an item as exhibiting DIF at a rate higher than a free-baseline model. In most instances, the gain of power obtained by using a constrained-baseline model versus a free-baseline model was overshadowed by the significant increase in FPs. Under conditions of large DIF, the increases of FP rates were drastic. For that reason, the *results suggest that a free-baseline model should be preferred*.

A third finding of the dissertation study is that the TP and FP rates also varied depending on the size and source of DIF. To be specific, *as the size of DIF increased, so did the TP and FP rates of each model given a constrained-baseline strategy. Given a free-baseline approach, increasing the size of DIF led to an increase of TP rates only.* The impact of increasing the size of DIF therefore was much greater for the constrained-baseline approach than the free-baseline approach.

The source of DIF also had an impact on the TP and FP rates of the constrained-baseline. Generally speaking, *the more item parameters simulated to exhibit DIF, the larger the TP and FP rates across all models given a constrained-baseline approach. Given a free-baseline approach, the source of DIF only impacted the TP rates of the studied models.*

Lastly, the results of the study suggest that, *given a constrained-baseline model, increasing the sample size will lead to an increase in both FP and TP rates* (Stark et al., 2006). As a matter of fact, given large DIF, the FP rates of the constrained-baseline models were well above the nominal rate of .05 (see Chapter 4). *Given a free-baseline model, increasing sample size led to an increase of the TP rates only.* These results provide additional support for the use of a free-baseline model.

5.2 Discussion of the Results

The purpose of this section is to discuss the results of the dissertation study as it relates to the literature at large. As previously mentioned, the GR IRT model using a free-baseline strategy is the preferred model of choice. These results are similar to the results found by Meade and Lautenschlager (2004) and Kim and Yoon (2011). Given that the

design of the Meade and Lautenschlager (2004) and Kim and Yoon (2011) studies varied greatly the results were not directly comparable. Because this dissertation study covered the design and methods used by both of these studies, it is now clear that their results were not due to issues in design.

A second finding of this study is that a free-baseline model is preferred over a constrained-baseline model. This finding is similar to the findings of Stark et al. (2006). Both studies found that using a constrained-baseline model versus a free-baseline model led to inflated FP rates. Because Stark et al. (2006) only compared the LCFA model and the GR IRT model, this study contributes to the literature by including the CCFA model. In addition, considering that Stark et al. only simulated 50 replicates per condition, this study, which simulated 1,000 replicated per condition, provides evidence that suggests that Stark et al.'s results were not due to chance.

As previously mentioned, the size and source of DIF also had an effect on the TP and/or FP rates of the studied models. In similar fashion, Kim and Yoon (2011), Meade and Lautenschlager (2004), and Stark et al. (2006) found that increasing the size of DIF and number of item parameters exhibiting DIF led to an increase in TP and/or FP rates. For instance, it is easier to correctly identify an item as exhibiting DIF if all of the threshold parameters exhibit DIF as compared to when only the discrimination parameter exhibits DIF.

Finally, the current study found that, generally speaking, increasing the sample size will lead to an increase in TP rates. Given a constrained-baseline model, increasing the sample size will also lead to an increase in FP rates. In contrast, given a free-baseline

model, an increase of sample size had little to no effect on FP rates. These results are similar to those found by Kim and Yoon (2011), Meade and Lautenschlager (2004) and Stark et al. (2006). It is important to note, however, that Meade and Lautenschlager (2004) used a free-baseline model, whereas Kim and Yoon (2011) used a constrained-baseline model. Therefore, the results of this study pertaining to the constrained-baseline model analyses are similar to those of Kim and Yoon (2011), whereas the results of this study pertaining to the free-baseline model analysis are similar to those of Meade and Lautenschlager (2004).

5.3 Significance of the Study

The results of this dissertation study are important for researchers throughout social science research. For instance, applied cross cultural researchers often develop measurement instruments that are developed to compare hypothesized constructs across cultural groups (Chen, 2008; Riordan & Vandenberg, 1994). Typically, instruments are created in English, and translated to specific languages of interest. Given the scores obtained from the researcher-designed measurement instrument, comparisons are made across cultural groups. In order for these comparisons to be valid, measurement invariance is an assumption that must be tenable (Gregorich, 2006; Millsap, 2011).

Assuming that the measurement instrument is composed of Likert-scaled items, the item responses are on an ordinal scale (Lubke & Muthén, 2004). Research suggests that the LCFA model is appropriate when there are at least five response options (Babakus et al., 1987; Muthén & Kaplan, 1985). It is important to note that Babakus et al.'s (1987) and Muthén and Kaplan's (1985) findings are based upon a single group. The

results of this study, however, suggest that the LCFA model is not appropriate when the item responses are on an ordinal scale given multiple groups (Lubke & Muthén, 2004; Temme, 2006). Thus, the significance of this dissertation also entails contributions made to the literature studying the appropriateness of the LCFA model under conditions of ordered-categorical data.

Based on the results from prior research (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006), applied researchers interested in assessing measurement invariance using measurement instruments composed of 5-point Likert scaled items may be confused as to the appropriate model to use. On one hand, Meade and Lautenshlager (2004) argue that the LCFA model has lower power than the GR IRT model in detecting DIF. On the other hand, Stark et al. (2006) conclude that the LCFA model has comparable power to the GR IRT model in detecting DIF. Furthering the complexity of the issue, Kim and Yoon's (2011) comparison of the CCFA model to the GR IRT model found that the GR IRT model is the preferred choice of model. Taking all three studies into consideration, an applied researcher may be unsure as to which model to use (i.e., LCFA, CCFA, or GR IRT models). Given the focus, models, and factors manipulated in this dissertation study, these issues (contradictions) have been addressed. Applied researchers now have guidance as to the most precise model to use when assessing measurement invariance with 5-point Likert scaled data. That is, the GR IRT model is the preferred model of choice, similar to the results of Kim and Yoon (2011) and Meade and Lautenschlager (2004), and that the free-baseline model is preferred.

The results of this study also have implications on the field of education as it relates to classroom-level assessments. Teacher made tests are likely to include biased items simply because teachers do not possess the skills needed to test for items that are exhibiting DIF (Popham, 2003, 2009). In addition, a classroom teacher may not teach enough students to use an unobserved conditional invariance technique for DIF detection. Popham (2003) argues that "a teacher who is instructing students from racial/ethnic groups other than the teacher's own racial/ethnic group might be wise to ask a colleague (or a parent) from those racial ethnic groups to serve as a one-person bias review committee (p.58)". While this approach is subjective in nature, it recognizes that teacher-made assessment may unintentionally include items that are biased. Considering that day-to-day instructional decisions are made based upon classroom-level assessment data, making an attempt to remove any bias from a classroom-level assessment is needed and worthwhile (Popham, 2009).

Instead of teachers creating assessments at the classroom level, it is argued that assessments should be created at the district-level. By creating assessments at the district level, it is likely that the sample size is large enough for a quantitative investigation of DIF using an unobserved conditional invariance technique. In addition, the district can employ education measurement specialist with high levels of assessment literacy that are familiar with quantitative and psychometric methods. The results from this study can contribute to education measurement specialist at the district level by providing guidance on the most accurate unobserved conditional invariance techniques for DIF detection

during the creation of district-level assessments. By creating fair and equitable tests, we are striving closer to ensuring a fair and equitable education for all students.

5.4 Limitations

In spite of the contributions made by the dissertation study, it is important to remember that no study is free of limitations. Limitations of this dissertation study include: (a) only a single dimension was simulated, (b) only a single item exhibited measurement non-invariance, (c) the number of response options was not manipulated, (d) the data were generated using a CCFA model, (e) only two groups were simulated, and (f) the sample sizes of the groups were balanced.

5.5 Recommendations for Future Work

Based on the results of the dissertation study, the research questions of interest have been clearly answered. Nonetheless, there is additional work that needs to be done to further the quantitative research literature. Additional research areas of interest include: (a) comparing unobserved and observed conditional invariance techniques for DIF detection (Elosua, 2011), (b) studying DIF detection given multidimensional data, (c) studying DIF given multiple groups of unbalanced sample sizes,(d) studying DIF detection across three or more groups, (e) studying DIF detection when multiple items on a scale exhibit measurement non-invariance (Stark et al., 2006), (f) studying DIF detection when the data are generated using an IRT model (Meade & Lautenschlager, 2004), (g) studying DIF detection given different baseline models and different parameterizations (Millsap & Yun-Tien, 2004), and (h) comparing the ability of fit

indices other than the χ^2 statistic (e.g., Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Weighted Root Mean Square Residual (WRMR); Kim & Yoon, 2011).

5.6 Final Thoughts

Within the literature comparing unobserved conditional invariance techniques to DIF detection, there were many inconsistencies across studies that led to contrasting results. For instance, Stark et al. (2006) found that the LCFA model is the preferred model for DIF detection, whereas Kim and Yoon (2011) and Meade and Lautenschlager (2004) found that the GR IRT model is the optimal model for DIF detection. The goal of the current dissertation study was to address some of those inconsistencies by integrating the features of the three most prominent studies (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Stark et al., 2006) within a single study. Towards that end, this study found that the IRT model with a free-baseline model is the optimal model when studying DIF using Likert-scaled data. As previously mentioned, there are limitation of this study that warrant further research (e.g., only a single item exhibited DIF). In spite of these limitations, the goal of the study was achieved and the research questions were answered. Given the results of the study, researchers studying DIF using polytomous data should select the GR IRT model and use a free-baseline model.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*(2), 222-228.

Baker, A. J. L, & Soden, L. (1997). *Parent involvement in children's education: A critical assessment of the knowledge base.* Paper presented at the annual meeting of the American Education Research Association, Chicago, IL. Retrieved from ERIC. (ED407127)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp 397-479). Reading, MA: Addison-Wesley

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. *Psychometrika*, *46*(4), 443–459.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248. DOI: 10.1037/a0023350

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018. DOI: 10.1037/a0013193

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5-32.

Davis-Kean, P. E. & Sexton, H. R. (2009). Race differences in parental influences on child achievement: Multiple pathways to success. *Merrill-Palmer Quarterly*, *55*(3), 285-318.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in Mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(4), 610-632. doi: 10.1080/10705511.2012.713272

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Desimone, L. (1999). Linking parent involvement with student achievement: Do race and income matter? *The Journal of Educational Research, 93*(1), 11-30.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 327-346. doi: 10.1207/S15328007SEM0903_2

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*(3), 217-23.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92*(2), 526-531.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86.

Edwards, J. E., Elig, T. W., Edwards, D. L., & Riemer, R. A. (1997, April). *The 1995 Armed Forces Sexual Harassment Survey: Codebook for Form B* (Report No. 95-014). Arlington, VA: Defense Manpower Data Center.

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicológica*, *32*(2), 403-421.

Finney, S. J. & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269-314). Greenwich, CT: Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491. doi: 10.1037/1082-989X.9.4.466

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, *23*(4), 309-326.

Forero, C. G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275-299. doi: 10.1037/a0015825

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11), S78-S94.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*, Boston, MA: Kluwer · Nijhoff.

Hagedorn, M., O'Donnell, K., Smith, S., & Mulligan, G. (2008). *National Household Education Surveys Program of 2007: Data File User's Manual, Volume III, Parent and Family Involvement in Education Survey.* (NCES 2009-024). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Hill, N. E. & Craft, S. A. (2003). Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families. *Journal of Educational Psychology, 95*(1), 74-83. doi: 10.1037/0022-0663.95.1.74

Holland, P. W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Kim, E. S. & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 212-228. doi: 10.1080/10705511.2011.557337

Long, J. S. (1983). *Confirmatory factor analysis*. (Sage University Paper series on Quantitative Application in the Social Sciences No. 33.) Beverly Hills, CA: Sage.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Love, Q. U. (2012). *Latent variable models in parental involvement research.* Unpublished manuscript.

Lubke, G. H. & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons, *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(4), 514-534.

MacCallum, R. C. (2009). Factor analysis. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 123-147). Thousand Oaks, CA: Sage.

MacIntosh, R. & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372-379. Doi: 10.1177/0146621603256021

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc

Meade, A. W. & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388. doi: 10.1177/1094428104268027

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016-1031. doi:10.1037/a0027934

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-534.

Meredith, W. & Teresi, J. A. (2006). A essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69-S77.

Millsap, R. E. (1998). Group differences in regression intercept: Implication for factorial invariance. *Multivariate Behavioral Research, 33*(3), 403-424.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York: Routledge

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Millsap, R. E. & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115. doi: 10.1037/1082-989X.9.1.93

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*(1), 3-31.

Mplus (Version 6.12). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomized variables. *Psychometrika*, *43*(4), 551-560.

Muthén, B. O. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*(1), 43-65.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Newbury Park, CA: Sage.

Muthén, B. O. (1998-2004). *Mplus technical appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407-419.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript. Retrieved from http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171-189.

Muthén, L. K. & Muthén, B. O. (1998-2010). *Mplus user's guide*. Sixth edition. Los Angeles, CA: Muthén & Muthén.

Oishis, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*, 411-423. doi: 10.1016/j.jrp.2005.02.002

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential function of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*(1), 1-17.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning*. Second Edition. (Sage University Paper series on Quantitative Application in the Social Sciences No. 161.) Thousand Oaks, CA: Sage.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models.* (Sage University Paper series on Quantitative Application in the Social Sciences No. 144.) Thousand Oaks, CA: Sage.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J. & Chen, F. (2001). Monte Carlo experiments: Design and Implementation. *Structural Equation Modeling*, *8*(2), 287-312.

Popham, W. J. (2003). *Test Better, Teach Better: The Instructional Role of Assessment*. Alexandria, Va: Association for Supervision and Curriculum Development.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*(4), 4-11. doi: 10.1080/00405840802577536

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517-529. doi: 10.1037//0021-9010.87.3.517

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353-368. doi: 10.1177/014662169501900405

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*(3), 643-671. DOI: 10.1177/014920639402000307

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(4, Pt.2), 1–100.

SAS® (Version 9.3). [Computer Software]. Cary, NC: SAS Institute Inc.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363. doi: 10.1177/0734282911406661

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292-1306. doi: 10.1037/0021-9010.91.6.1292

Swaminathan, H. and Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Temme, D. (2006). Assessing measurement invariance of ordinal indicators in cross-national research. In S. Diehl & R. Terlutter (Eds.), *International advertising and communication: Current insights and empirical findings* (pp. 455-472). Wiesbaden, Germany: Deutscher Universitäts-Verlag | GWV.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Retrieved from http://www.unc.edu/~dthissen/dl.html

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58-79. doi:10.1037/1082-989X.12.1.58.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*, *32*(7), 511-526. doi: 10.1177/0146621607310402

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi:10.1177/0146621607314044

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparisons to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1-27. doi: 10.1080/00273170802620121